DOCUMENT RESUME

ED 150 108 95 SP 011 860

AUTHOR Shavelson, Richard; Dempsey, Nancy
TITLE Generalizability of Measures of Teacher Effectiveness and Teaching Process. Beginning Teacher Evaluation Study, Technical Report Series No. 75-4-2.
INSTITUTION Far West Lab. for Educational Research and Development, San Francisco, Calif.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C..
PUB DATE [75]
CONTRACT 400-75-0001
NOTE 109p.

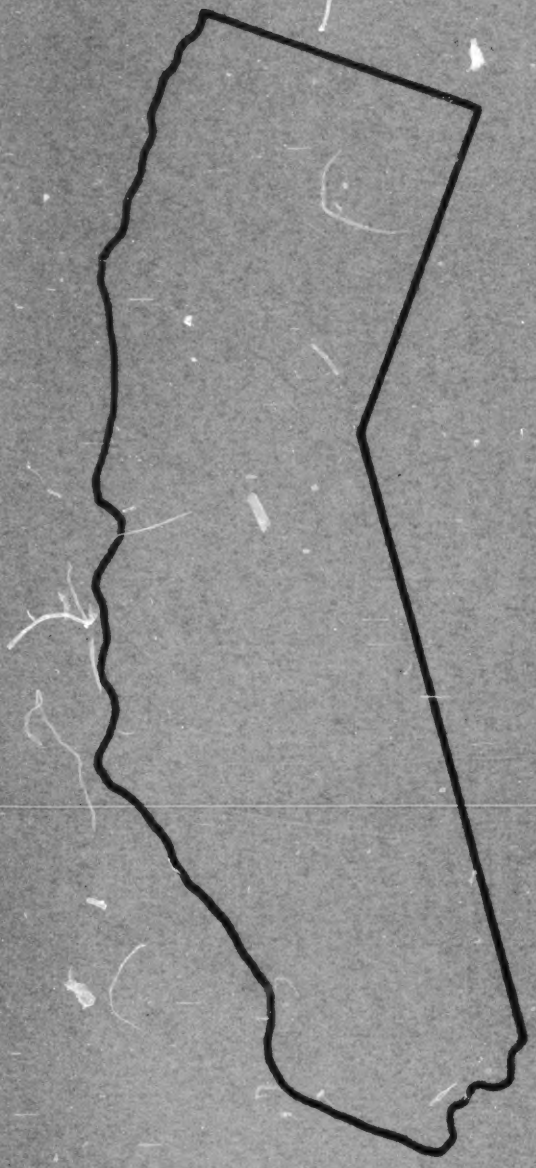
EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Academic Achievement; Class Management; Educational Research; *Effective Teaching; *Interaction Process Analysis; *Measurement Techniques; Student Attitudes; Student Motivation; Teacher Attitudes; *Teacher Behavior; *Teacher Evaluation; Teaching Methods

ABSTRACT

This report reviews studies of the stability of measure of teacher effectiveness and behavior. In studying teacher effectiveness over short intervals it appears that teaching the same content to similar students yields moderately stable generalizability coefficients over occasions. However, when teachers teach different content to the same or similar students, the measured effectiveness is not very stable over occasions. Studies of the stability of measured effectiveness over long time intervals yields very low to moderate stability coefficients. When the stability of the behavior of teachers was examined, global ratings appeared to show greater teacher stability over occasions than did frequency counts. Nevertheless, in general, there was very little evidence found to indicate that measures of teacher behavior over time are stable. Instability of teacher behavior seemed to be the common finding. Implications for research on teaching are presented. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED150108



BTES

BEGINNING TEACHER EVALUATION STUDY

Technical Report Series

Technical Report #75-4-2

Generalizability of Measures of Teacher Effectiveness and Teaching Process

Richard Shavelson and
Nancy Dempsey

University of California,
Los Angeles

Best copy available

SP 011 860

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

FAR WEST LABORATORY
FOR EDUCATIONAL RESEARCH AND DEVELOPMENT
1855 FOLSOM STREET • SAN FRANCISCO, CALIFORNIA 94103 • (415) 565-3000

This work was completed pursuant to Contract #400-75-0001 between the Department of Health, Education and Welfare, National Institute of Education and the California Commission for Teacher Preparation and Licensing.

The Far West Laboratory for Educational Research and Development, 1855 Folsom Street, San Francisco, California 94103, is a nonprofit organization supported in part by the United States Office of Education and the National Institute of Education, Department of Health, Education and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of these agencies, and no official endorsement by them should be inferred.

Generalizability of Measures of Teacher Effectiveness and Teaching Process*

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: SP TM
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

Richard Shavelson

and

Nancy Dempsey

University of California,

Los Angeles

*We wish to express our thanks to David Berliner, Barak Rosenshine, and Philip Winne for their critical review of this paper.

GENERALIZABILITY OF MEASURES OF TEACHER EFFECTIVENESS AND TEACHING PROCESS

Richard Shavelson and Nancy Dempsey

University of California, Los Angeles

Recent research on teaching has followed two lines of inquiry. One is the criterion of effectiveness paradigm and the other is the teaching process paradigm (Gage, 1972). The criterion of effectiveness paradigm uses pupil outcomes--usually achievement--as a measure of teacher effectiveness. In contrast, the teaching process paradigm focuses on various aspects of teacher and student behavior judged to be worthwhile in their own right or linked to student achievement in previous research.

Neither paradigm has identified features of human teaching which lead directly--or even indirectly--to valued student outcomes. Consistent conclusions from research on teaching are that teacher effects on pupil outcomes are unstable (Rosenshine, 1970), that teaching acts themselves may be unstable (e.g., Moon, 1971), and that most teaching acts are unrelated to student outcomes (Heath & Nielson, 1973; Rosenshine & Furst, 1971).

One possible explanation for these conclusions is that typical measures of teacher effectiveness (e.g., students' mean achievement) are unstable and measures of teacher behavior (e.g., teachers' use of verbal reinforcements) may be too unstable to yield consistent relationships with student outcomes (cf. Berliner & Ward, 1974). This observed instability over time may be due to one or some combination of the following: (a) systematic and/or random changes in teaching behavior over time, (b) systematic and/or random differences in teaching from one

situation to the next, (c) systematic differences between pupils taught from one time to the next, and (d) limitations in measurement methods. Regardless of the sources of instability over time, one implication of this position is that alternative conceptualizations of measures of teacher effectiveness and behavior may need to be considered. Such alternative conceptualizations might include teacher effectiveness, measures in standardized test situations (cf. McNeil & Popham, 1973), or teacher process measures of variability or flexibility of teaching acts (cf. Flanders & Simon, 1970).

Another possible explanation is that the stability of behavioral measurements used in most research has not been adequately examined or established (cf. Medley & Mitzel, 1963). For example, in most studies of, say, teaching process, a random sample of teachers is observed by two or more raters. The consistency with which the raters rank order teachers on some variable such as "number of verbal reinforcements" is interpreted as the reliability of the measurement, i.e., interrater reliability. Typically, each teacher's score on this variable is an average of the raters' scores for that teacher. While not stated explicitly, this score may be considered characteristic of the teacher's use of verbal reinforcements regardless of the nature of the students taught, the instructional goals, the teaching situation, and so on. In short, while the measurement is taken in one particular setting at a particular point in time, the measurement is interpreted as generalized over many settings at different points in time. A prime example of generalization are studies in which observations of teachers' verbal reinforcements obtained at some point in time are correlated with student achievement at the end of a course. While it is possible to

consistently rank order teacher performance at one point in time (interrater reliability), it is an empirical question as to whether this rank ordering is stable (generalizes) over teaching settings. It may have been that, by chance, a particular teacher was teaching a lesson which presented numerous opportunities to reinforce students. This teacher would probably earn a high score on verbal reinforcements. If observed again, a different lesson might entail considerable explanation and permit few opportunities for reinforcement. The same teacher would probably earn a low score this time.

This analysis suggests that, in addition to determining the generalizability of scores over raters (i.e., interrater reliability), the consistency of rank orderings of teachers at different points of time teaching the same or different lessons to the same or different students (etc.) should be examined and the limits of generalizability established. Thus, commonly used measures of teacher behavior and effectiveness should not be abandoned until problems of the generalizability of the measurements have been thoroughly examined. A resolution of these measurement problems would then provide a basis for determining whether to abandon commonly used measures of teacher effectiveness and behavior.

This paper re-examines studies of the stability of measures of teacher effectiveness and behavior. In attempting to resolve the measurement problem, it would be ideal to find studies in the literature which systematically varied such facets[1] of their measurements as raters, occasions, subject matter, and student abilities in order to determine the effect of these facets on a teacher's behavior. Unfortunately this is seldom the case (for exceptions, see Fortune, 1966a, b, 1967; Justiz,

1969; Millman, 1973; Marzano, 1974; Sandoval, 1974). The next best alternative is to categorize studies according to characteristic facets such as teachers and students, grade level, occasions, subject matters, and so on in order to examine the external validity (cf. Bracht & Glass, 1968; Campbell & Stanley, 1963; Snow, 1974) or generalizability of stability coefficients from studies which are similar and/or different on a variety of facets. In so doing, facets which limit the generalizability of findings from one study to the next might be considered prime candidates for inclusion in a study to determine whether the facet affects the same teacher's teaching in the same way--i.e., limits the generalizability of the measurement. We entertain the possibility that generalizability may be extremely limited in an educational context (cf. Glass, 1972b) and that this possibility, along with the design of existing research, may place severe limitations on our attempt to examine the measurement problem in research on teaching through a review of literature.

Studies of Teacher Effectiveness

The criterion of effectiveness paradigm uses student outcome measures (e.g., achievement, attitude) as the criterion for identifying effective teachers. Within this paradigm there are a variety of definitions of teacher effectiveness. For example, Gage (1963, p. 116) defined teacher effectiveness as "the teacher's effect on the realization of some value. Usually, the value takes the form of some educational objective, defined in terms of desired pupil behaviors, abilities, habits, or characteristics." Scriven (1974) noted that good teaching must not only make significant contributions to meeting the educational needs of the student and society but must also be reasonably effective compared to available teaching

alternatives.

Other definitions present more restricted views of teacher effectiveness in strongly recommending only student outcome measures to evaluate teachers. For example, McNeil and Popham (1973, p. 220) argued that "the test of a teacher's effectiveness is his accomplishment of the goals of education as displayed in the pupil's behavior following instruction." Rosenshine (1970, p. 647) defined teacher effectiveness as the "residual class mean achievement scores in which a measure of prior achievement or student aptitude is used to adjust posttest scores by regression."

None of these definitions explicitly include or exclude the notion that effectiveness is a generalizable characteristic of a teacher. Nor do these definitions specifically delineate the conditions under which effectiveness should be demonstrated. But it is possible that certain teaching behaviors are effective with certain groups of students while others are ineffective with these students. Since teachers typically use one or another set of teaching methods, teacher effectiveness as currently measured may depend on the luck of matching a teacher with a particular group of students. In addition, a teacher who happens to be effective in the situation in which he or she is fortuitously placed may not be effective in another situation. This implies that measures of teacher effectiveness may lack generalizability--i.e., may be unstable when collected with different students and in different settings.

Data on the generalizability of measures of teacher effectiveness are reviewed below to examine this possibility. Since teaching settings vary in many ways, it is important to characterize the studies as to the various settings or facets under which the measurements of teacher

effectiveness were made. The ideal case would be one in which, for a particular population of teachers, various situational facets (e.g., student ability, subject matter) and observational facets (e.g., criterion- or norm-referenced measures, number of measurement occasions) are systematically varied. Then statements about the influence of a particular facet on the consistency of teacher effectiveness measures could be made.

But most of the studies were not designed to systematically vary observational and situational facets. In order to examine the influence of a particular facet or set of facets on the generalizability of measurements, then, data from several studies--each study representing a level or condition of one or more facets--must be considered for the purposes of this review. In examining certain facets over different studies, other facets cannot be held constant. Thus, conclusions about the generalizability of measures over certain facets will be confounded by unsystematic variation in other facets of the studies being examined. At best, data from a set of studies can be examined for patterns suggesting the limits of generalizability. In this way, facets which may influence measures of teacher effectiveness can be identified for further study.

Teacher effectiveness studies can be divided into long-term investigations in which instruction usually lasted a full school year or longer and short-term investigations in which instruction consisted of lessons of 30 minutes in duration or less (cf. Rosenshine, 1970). Rosenshine (1970) provided the most complete review of these studies to date. In the studies he reviewed, the primary index of the generalizability (stability) of measures of teacher effects were correlations between mean residual gain scores obtained at two or more points in time.

Rosenshine (1970) located five long-term studies of teacher effectiveness. Those studies, and additional studies by Brophy (1973), Creemer (1974), and Veldman and Brophy (1974) are outlined in Table 1.

- -

Insert Table 1 about here

- -

The rows of the table identify the studies. The columns provide information about them, including various situational and observational facets. (These facets are representative, to a greater or lesser extent, in the studies reviewed here and in other sections of the paper.)

The entries in the table provide critical information about each study. Entries under the student, content, teacher, and dependent-variable categories may be interpreted as follows. The column labeled assignment identifies how students were assigned to teachers: randomly, quasi-randomly (students within an intact classroom were randomly assigned to one of two teachers; more than two teachers or classrooms were used in the study), or nonrandomly. The similarity facet identifies the similarity of students taught across observation occasions: same (exactly the same students at two or more occasions), similar (students drawn from the same population as in the first observation), or different (students from different populations at different occasions). The grade level and type of student facets are self-explanatory.

Two content facets were identified. The type facet refers to the specific subject matter taught during the study period. The similarity facet refers to the correspondence of curricula from one observation to the next: same (exactly the same topics or curricula), similar (same general content area but perhaps differing in specific topics, curricula,

or in an unspecified way), or different (topics or curricula differing in content and perhaps in other unspecified ways).

Teachers in each study are identified as practicing (P) if they were employed as a regular teacher in a school, trainee (T) if they were enrolled in a teacher training program, or undergraduate (U) if they were students enrolled in a school of education but not necessarily pursuing a teaching credential at the time of the investigation.

Finally, important facets of the dependent variable are identified. One facet indicates whether posttest questions were provided to teachers in an attempt to increase the congruence between instruction and evaluation. Two other facets identify the type of meaure (NR=norm-referenced, OR=objectives-referenced) and content of the test. Finally, stability coefficients are reported for each study.

In the study by Morsh, Burgess, and Smith (1955), 106 Air Force instructors taught a short course (eight daily, one-hour sessions) on airplane hydraulics to two different classes one month apart. Residual gain scores on a written posttest and on a performance test were calculated. Pretest scores and grades in previous courses in the training program were used as covariates. Mean residual gain scores for the first class were correlated with similar scores for the second class. The stability coefficients were 0.34 for the written test, 0.32 for the performance test, and 0.38 for the two tests combined.

In the Harris, Morrison, Serwer, and Gold (1968a, b) studies, posttest scores on published achievement measures of reading and language skills were obtained for a sample of 30 first-grade and a sample of 24 second-grade teachers who taught similar classes in two successive years. Posttest scores were adjusted using five reading and language

pretest scores as covariates. In the first-grade sample, stability coefficients ranged from 0.52 to 0.53 for spelling and word reading, from 0.32 to 0.36 for paragraph meaning and word study skills, while the stability coefficient for the vocabulary subtest was -0.01. The correlations for the second-grade sample indicated much less stability, ranging from -0.08 to 0.26 for the same measures.

Soar (1966) investigated teacher and student classroom behavior and student achievement in 55 classrooms, grades three through six, for two consecutive years. Residual gain scores were calculated for each student using pretest and posttest scores on the following four subtests from the Iowa Tests of Basic Skills: vocabulary, reading, arithmetic concepts, and arithmetic problems. A factor analysis of both teacher behavior indices and student outcomes revealed a factor labeled "achievement gain." It contained high loadings for pupil gain on the four achievement measures. Correlations between teacher scores on this factor was 0.09 over two years.

In the Torrance and Parent (1966) study, only descriptive statements about teacher consistency were reported. Sixty-three teachers using School Mathematics Study Group materials in classes of seventh through twelfth graders were studied for two consecutive years. Residual gain scores using the Sequential Tests of Educational Programs (mathematics) were calculated for each teacher. Teachers falling in the upper and lower thirds of the distribution of residual mean scores were identified for further study. Correlations between teacher effects across the two-year period were not reported. However, according to anecdotal report, there was little consistency between those in the "most effective" and those in the "least effective" criterion groups.

Brophy (1973) compared student achievement on the Metropolitan Achievement Test of reading and arithmetic in five Title I and 35 non-Title I schools. Second-and third-grade classrooms were studied over a three-year period. Only teachers who taught the same grade for three consecutive years were included in the sample. Residual gain scores were calculated using grade-equivalent scores collected at the beginning of the school year as covariates and the same measures collected in the fall of the next year as the dependent variable.

Stability coefficients for mean residual gain scores over the three years and median coefficients over the three-year period for grade level and type of student population were reported. The median stability coefficients for grade level and type of student population are presented in Table 1.

While the median stability coefficients reflect the general pattern of the data for all samples except the Grade 2, Title I sample, they must be cautiously interpreted. For example, while the median coefficient for the Grade 2, non-Title I sample was 0.42, the stability coefficients for the arithmetic concepts subtest over the three-year observation period were 0.34, 0.45, and 0.06. Similarly, although the median coefficient for the Grade 3, non-Title I sample was 0.40, a much less stable pattern emerged for the reading subtest (0.26, -0.07, 0.10). Inconsistent patterns were also obtained for the following subtests for the Grade 3, Title I sample (median coefficient=0.39): word discrimination (0.28, 0.19, 0.02), reading (0.54, 0.39, 0.08), and arithmetic reasoning (0.28, 0.32, -0.19).

In the Grade 2, Title I sample, the pattern is more complex. Data for years one and two indicate a similar pattern of stability, although

somewhat lower in magnitude as found in the Grade 2, non-Title I schools. However, teacher effects are highly unstable when observations from year one or year two are compared with year three. For example, the median stability coefficient is 0.09 for years two and three, and 0.06 for years one and three. The cause of this instability is not clear from these data. It may be that students in year three differed systematically in some way from students in years one and two.

Brophy (1973) also presented an alternative way of viewing consistency which is noteworthy. The residual mean gain scores for each teacher on each subtest over the three-year period were examined for one of the following patterns: (1) linear consistency where a teacher showed similar residual mean gains each year for a particular subtest (e.g., 0.27, 0.31, 0.29); (2) linear trend where a teacher showed improvement in residual means each year (e.g., 0.03, 0.24, 0.42) or the teacher showed a decrease each year (e.g., 0.38, 0.16, -0.08); and (3) nonlinearity where a teacher showed an inconsistent pattern of residual means over time.

In Brophy's sample, 28 percent of the teachers showed linear consistency and 13 percent showed linear improvement, while 11 percent showed linear decline and 49 percent showed nonlinearity. This method provides an interesting approach for examining teacher consistency. However, designations of teachers to these categories was based on the author's judgment, which was not described. If this approach is to be useful in the future, a more systematic procedure for identifying linearity or nonlinearity is required.

Veldman and Brophy (1974) re-analyzed the data from the Brophy (1973) study with a more restricted sample. Teachers selected from the

Brophy (1973) sample for inclusion in this study were those who had at least (a) five years of teaching experience at their grade level and (b) 14 students with complete data for each of the three years of the study. Intraclass correlations were calculated with mean residual gain scores averaged over the three years for each of the four samples of teachers (see Table 1). The stability coefficients for each outcome measure for the Grade 2, non-Title I sample are fairly high, ranging from 0.48 to 0.74. The coefficients for the Grade 3, non-Title I sample are also fairly high, ranging from 0.38 to 0.80 with one exception (0.23). Similarly, the coefficients for the Grade 3, Title I sample are also fairly high, ranging from 0.50 to 0.78 with two exceptions (0.26, 0.27). The Grade 2, Title I sample showed the least stability with coefficients ranging from 0.00 to 0.43.

In general, the pattern of stability coefficients resembles the pattern of median coefficients for these samples reported by Brophy (1973) in which the Grade 2, Title I sample showed the least stability and the other three samples showed somewhat more stability. However, the correlations are generally higher in the Veldman and Brophy study than the earlier Brophy study. Since the major difference between the teachers in the Veldman and Brophy study and the Brophy study was years of experience at a particular grade level, this increase in stability may be due, in part, to the greater experience of the teachers in the Veldman and Brophy study.

Creemers (1974; & Weeda, 1974) investigated teacher effectiveness in teaching initial reading. Both norm-referenced and objectives-referenced measures of student achievement were used. Details of the study and data are not available in English at the present time. However, the

author concludes that a relationship exists between teaching styles and pupil achievement. Furthermore, he states that "such relationships generally do not appear to be different among groups of pupils with different initial characteristics."

Summary: Long-Term Studies of Teacher Effectiveness

In discussing the first five studies in Table 1, Rosenshine (1970) pointed out that "Five studies hardly form a sufficient basis for generalization, even though the studies were widely discrepant in settings, subject areas, and students" (p. 649). "(O)n the basis of these studies, evidence on the consistency of teacher effects is weak because correlations as high as 0.5 were obtained in only one study (Harris et al., first grade, 1968), and all other correlations were about 0.35 or much lower" (pp. 650-651). The inclusion of three additional studies (a 60 percent increase) does not change the tenor of Rosenshine's conclusions. While Brophy (1973) and Veldman and Brophy (1974) report consistently higher coefficients than reviewed by Rosenshine, the difference is due, in part, to the method of calculating coefficients. Brophy and Veldman, by averaging coefficients over three years of observations, would be expected to report higher coefficients than studies reporting individual coefficients.

The most striking finding in Table 1 is that the stability coefficients vary unsystematically in magnitude. While our original intent was to examine each facet individually, lack of covariation between stability coefficients and facets makes this a tedious and unrewarding task. For example, in the Harris et al. (1968) study, the stability of teacher effects on spelling in grade one was 0.52 and in grade two it was 0.21, while for vocabulary the stability in grade one was -0.01 and in grade

two it was 0.19. In the Brophy (1973) study, the stability of teacher effects on reading scores ranged from 0.54 for data from years one and two to 0.08 for data from years one and three (Grade 3, Title I sample). Finally, within a particular grade level with students drawn from a particular population studying particular curricula, the stability of effectiveness measures varied greatly from one outcome measure to the next, even when these measures were similar. For example, in the Harris et al. (1st grade, 1968) study the stability of teacher effects on vocabulary was -0.01 while it was 0.32 for paragraph meaning and 0.52 for spelling.

Rosenshine noted a number of threats to the internal validity of these studies such as statistical adjustment for equating non-random groups, appropriateness of the criterion tests (norm-referenced versus criterion-referenced), mortality, testing, and so on. Since these threats are well known in the literature (Campbell & Stanley, 1963), they will not be discussed here; the reader is referred to Rosenshine (1970) for their application to these studies.

While threats to the internal validity of these studies must be a first concern, the focus of this review points to the external validity of these studies as an equally critical concern (cf. Bracht & Glass, 1968; Snow, 1974). Snow (1974) has discussed three dimensions of external validity. One dimension, population validity, refers to the generalization from an observed sample to an accessible population to "the substantive phenomena under study" (Snow, p. 270). Since few studies adequately satisfy the steps in this chain of generalization, "the only recourse is

to detailed description of the target population, the accessible population, and sample at hand . . ." (Snow, 1974, p. 270). Using this criterion, all of the studies reviewed above lack population validity. Descriptions of the sample of teachers are usually inadequate as are the descriptions of students in the sample.

For example, Harris (1968a) provided the most detailed descriptions available. Teachers in this study were volunteers who varied greatly in age and experience. Their first-grade teaching experience ranged from 0-29 years. The mode was one year and the median was two years. Most had bachelor's degrees with some graduate work short of a master's degree. Students were first-grade Negroes from urban, low-income areas enrolled in schools which had scored at least one-half year behind the most recent citywide reading test in third grade. While these descriptions are more precise than most, they do not provide a sufficient basis for generalization.

A second dimension, ecological validity, concerns the extent to which the situations observed in the study are representative of the situations to which generalizations are to be made. Formally, a universe of situations should be defined and a random sample selected for study. Failing to satisfy this formal requirement, studies deal with accessible samples. As with population validity, the only recourse is an extensive description of the situations involved in the study. Again, the studies reviewed do not generally provide adequate descriptions of classrooms, curricula, etc., either at a molar or molecular level.

For example, Veldman and Brophy (1974) examined class mean achievement in reading and mathematics for teachers over a three-year period. However, the elements of the curricula (the molar level) are not discussed

and the adaptation of the curriculum by a particular teacher in his or her day-by-day teaching (the molecular level) was not observed. Therefore, when differences in stability coefficients are found for different populations on reading scores, it is not clear whether these differences represent different curricula, different emphases by teachers, instability of measures, or some other factors. The Harris et al. (1968) study is also a case in point. It was conducted in conjunction with two new reading curricula, a skills-centered approach, and a language experience approach. However, while the elements of the two curricula (the molar level) and the adaptation of the curriculum by a particular teacher in his or her day-to-day teaching (measured by daily logs) were discussed, data for teachers using the two different curricula were combined. Therefore, the source of differences in stability coefficients for, say, vocabulary and reading subtests, cannot be assessed.

The third dimension, referent generality, designates "the range or pervasiveness of possible experimental outcomes measured in a given study" (Snow, 1974, p. 273). This dimension represents Rosenshine's concern about the correspondence between instructional objectives and norm-referenced, outcome measures. But it goes beyond this. The effect of teaching on student outcomes is multivariate. Thus, attitudinal, aptitude, and personality measures should be considered in addition to achievement measures. Also, the outcomes measured should sample the range of effects from those most proximal to instruction (e.g., in the cognitive domain, criterion-referenced measures of acquisition) to those intermediate to the effects of instruction (e.g., norm-referenced tests) to those distal to the instructional effects (e.g., aptitude). Only in the Creemers (1974) study was referent generality systematically examined,

though only to a limited extent.

The long-term studies of teacher effectiveness, then, show low to moderate stability coefficients; however, an explanation of the variation in these coefficients does not emerge when various facets of the measurements are examined. In order to explain what appears, at present, to be unsystematic variation in stability coefficients, issues of internal and external validity of these studies must be addressed. From the generalizability perspective, precise identification of facets must await studies which more adequately address questions of population validity, ecological validity, and referent generality than the eight studies in Table 1.

Short-Term Studies of Teacher Effectiveness

Short-term studies of teacher effectiveness reviewed by Rosenshine and others by Conner (1969) and Millman (1973) are outlined in Table 2. Both rows and columns are arranged as in Table 1. Entries are made using the same assignment rules as described for Table 1. As with the long-term studies, the short-term studies will be described and their various conditions examined for patterns bearing on the stability of teacher effectiveness measurements.

- -

Insert Table 2 about here

- -

In the Fortune (1966a, b, 1967) experiments, two facets were systematically manipulated, similarity of students and similarity of content. These manipulations included: (1) teaching the same topics to similar students, (2) teaching different topics to the same students, or (3) teaching different topics to similar students. Each teacher taught

four lessons; however, the experimental design allowed six correlation coefficients to be calculated. This procedure yielded one stability coefficient for the first condition (Topic A taught to Group 1 and Group 2), two coefficients for the second condition (Topic A and Topic B taught to Group 1 or Topic A and Topic C taught to Group 2), and three coefficients for the third condition (Topic A taught to Group 1 and Topic C to Group 2 or Topic A to Group 2 and Topic B to Group 1 or Topic B to Group 1 and Topic C to Group 2).

In the first study (Fortune, 1966a), 30 randomly selected teacher trainees taught four 30-minute lessons on either transportation (e.g., truck, sailboat) or community workers (e.g., doctor, mailman) to groups of six to nine Headstart children. Assignment of students to teacher was quasi-random. Two lessons were presented by the teachers and two were presented on videotape. Posttest scores on an objectives-referenced measure were adjusted using learner group differences and difficulty of the particular lesson as covariates. Data for groups taught the transportation and community workers lessons were analyzed separately (Table 2).

Fortune, Gage, and Shutes (1966) used the same regression approach to obtain posttest scores on an objectives-referenced measure of achievement. Teacher trainees taught four social studies lessons to groups of five junior high school students. These students were randomly assigned to groups. All lessons were teacher-presented. An attempt was made to provide congruence between instructional objectives and test items by providing the teacher with five of the ten posttest questions beforehand. Stability coefficients for each condition are shown in Table 2.

The third study (Fortune, 1967) was similar to the first two except that undergraduates taught either four English, mathematics, or social

studies lessons to groups of fourth, fifth, or sixth graders in a university-related school. Data were analyzed separately for each subject matter (see Table 2).

In the Belgard, Rosenshine, and Gage (1968) study, 43 experienced volunteer teachers taught two, 15-minute social studies lessons to their own intact classes of high school students. Different topics were taught to exactly the same students. The stability coefficient was 0.47.

Justiz (1969) examined consistency of teacher effectiveness when different topics were taught to the same students or when different topics were taught to similar students. Student teachers taught two lessons, one on news story structure and one on computer punch cards to a randomly selected group of high school students. In one school these two topics were taught by seven student teachers to exactly the same group of students. The Spearman rank-order correlation based on rankings of student teachers on pupil group means for the two objectives-based tests was 0.90 for this sample. In a second school ten student teachers taught these two topics to two similar groups of high school students. The Spearman rank-order correlation for this sample was 0.63. However, these correlation coefficients must be cautiously interpreted. In the first school, two subjects were dropped from the study while one subject was dropped from the second school. This effect is difficult to assess, but missing cases can influence the results with such small samples (cf. Glass, 1972a).

In a similar study Connor (1969) assigned five high school students to experienced male teachers in a "nearly random" fashion. (Assignment was random with the exception that no teacher was assigned a pupil with

whom he was acquainted.) The teachers taught two different lessons (propaganda techniques and probability) to two similar groups of students. Connor reported a coefficient "corrected for attenuation" of 0.72. However, as Glass (1972a) pointed out, this is a highly inappropriate significance test. When either a Pearson product-moment correlation or a Spearman rank-order correlation is computed, both show less stability (r_{xy} =0.26, r_s =0.35).

In Study 1 reported by Millman (1973), stratified random sampling (by grade and academic ability) was used to form groups of fourth, fifth, and sixth graders from an urban school composed predominately by minority students. Each teacher taught three lessons on three different topics (light rays, adjectives, and folkways). The first two lessons were taught to the same group of children. The third lesson was taught to similar students drawn from the same population. A cognitive criterion test consisted of items corresponding to the instructional objective and similar to the sample questions provided the student teacher. An affective criterion test asked for each student's opinion of the lesson and the teacher. These tests were administered after each lesson. Teacher effectiveness scores were calculated for each student teacher using an analysis of covariance to adjust mean scores of students for initial differences on certain control measures. Teacher effectiveness scores were correlated over occasions. When teachers taught <u>different topics</u> to the <u>same</u> or <u>similar students</u>, the stability coefficients for the cognitive criterion tests were extremely low (0.25 to -0.03). The coefficients for the affective criterion tests under both these conditions were moderately stable (0.31 to 0.34).

In Study 2 the <u>similarity of students</u> and <u>similarity of content</u>

facets were systematically varied as was done in the Fortune (1966a, b, 1967) studies. Second and third graders drawn from a population similar to the one described in Study 1 were assigned to groups using stratified random sampling. Each student teacher taught lessons on decoding or rhythms on three occasions. Both cognitive and affective criterion tests were used to assess teacher effectiveness.

The most stable teacher effects as measured by the cognitive criterion test occurred when teachers taught different topics to the same students (0.71). Teacher effects were very unstable when the same topics were taught to similar students (-0.23) or when different topics were taught to similar students (0.09).

Greatest stability on the affective criterion test occurred when the same topics were taught to similar students (0.26). Less stability was observed when different topics were taught to the same students (0.18) and when different topics were taught to similar students (-0.22).

Summary: Short-Term Studies of Teacher Effectiveness

In reviewing the results of these short-term studies, the effect of the two facets systematically manipulated in a number of studies--similarity of students and similarity of content--is considered first. The Fortune (1966a, b, 1967) and Millman (1973) studies provide data on the stability of teacher effects when teaching the same content to similar students. The stability coefficients based on cognitive measures ranged from -0.23 (Millman) to 0.70 (Fortune, 1967). Millman (1973) reported a coefficient of 0.26 for an affective measure. Except for a stability coefficient of 0.22, all of those reported in the Fortune

studies were positive and moderately high. The Millman coefficient for the cognitive outcome, then, is the one discrepancy. However, Millman aggregated data for teacher trainees teaching rhythms lessons with data for trainees teaching decoding lessons. If the trainees benefitted more (learned to teach better) from teaching one kind of lesson (say rhythms) than the other, then the rank ordering of teachers on the second lesson would change considerably from the ranking on the first lesson. Hence, a zero or even negative correlation might be expected. Since Millman did not report mean scores for trainees in the rhythms and decoding groups separately, this reasoning cannot be empirically validated. However, this reasoning raises doubts about the Millman coefficient serious enough for us to ignore it for the present. We conclude, then, that the effects of teachers teaching the same lesson to similar students are moderately stable. This conclusion is consistent with the one long-term study of teacher effectiveness in which similar students were taught the same content (Morsh, et al., 1955); the stability coefficients in this study were positive and slightly lower (0.35) than most of those reported by Fortune (Rosenshine, 1970, p. 656, notwithstanding).

The effectiveness of teachers teaching different content to the same students was examined by Fortune (1966a, b, 1967), Belgard, et al. (1968), Justiz (1969), and Millman (1973). The stability coefficients varied from -0.43 (Fortune, 1967) to 0.90 (Justiz, 1969). If the Fortune (1967) study is ignored, the coefficients vary from approximately zero to 0.90 with one exception (-0.27, Fortune, 1966a).

The stability coefficients covary unsystematically with the facets identified in Table 2. For example, Rosenshine (1970) suggested that experienced teachers might be more stable in their effects than teachers

in training. But in correcting Rosenshine's categorization of one of the Justiz (1969) studies (from similar students/different content to same students/different content) and adding the Millman data, it is apparent that trainees can produce effects just as stable as those produced by experienced teachers. Stability coefficients for teachers teaching different lessons to the same students, then, are much less consistent and, surprisingly, greater in absolute magnitude than stability coefficients for similar students and same lessons. Since none of the long-term studies reported data on the same students being taught different lessons, no comparison can be made.

Finally, the effectiveness of teachers teaching different content to similar students was examined by Fortune (1966a, b, 1967), Justiz (1969), Connor (1969), and Millman (1973). The stability coefficients for achievement data ranged from -0.45 (Fortune, 1967) to 0.82 (Fortune, 1967). Except for the Fortune (1967) study, these coefficients ranged from approximately zero (Millman, 1973; Fortune, et al., 1966) to 0.63 (Justiz, 1969). For data on an affective measure, Millman (1973) reported stability coefficients of 0.31 (Study 1) and -0.22 (Study 2). As with the data for different content/same students, there is no consistency in the magnitude of the coefficients found using different content with similar students. The absolute magnitude of some of the coefficients (e.g., 0.82) is surprising since both content and students varied, a condition where one would expect the stability coefficients to be fairly low.

From the limited data available, the effects of teachers teaching the same content to similar students appears to be moderately stable and consistent over studies. The magnitude of stability coefficients for

teachers teaching different content to the same or similar students varies greatly. When the effects of other facets are examined in an attempt to explain this inconsistency, the variability in stability coefficients appears to be independent of the variability in facets.

While the discussion of internal and external validity above applies to the short-term studies reported here, one additional point should be made about the external validity, or more specifically the ecological validity, of the short-term studies. One aspect of ecological validity is the preparation of "tuning" of subjects: "The representativeness of an experiment (study) is enhanced if the learner's behavior is 'tuned,' that is, adapted to the task so that it is smooth and habitual as in ordinary school learning. Unnatural treatments--and any newly invented instructional method will be unnatural in terms of student past experience--require adaptation of the learner to be properly evaluated" (Snow, 1974, p. 280). In most of the short-term studies the content taught was selected by the experimenter and was unfamiliar to the teachers. Teachers were also unacquainted with the students who were randomly assigned to them. Similarly, topics were deliberately selected which were not part of the ordinary curriculum so that they would be unfamiliar to the students. This is not typical of a natural teaching situation. Therefore, both teacher behavior (i.e., teaching performance) and learner behavior (i.e., achievement, attitude) in these studies may reflect lack of adapatation to the task and may not be representative of ordinary school performance.

Studies of Teacher Behavior

The teaching process paradigm focuses on teacher and student behaviors

identified as important in their own right or linked to desirable student outcomes in previous studies. A wide variety of teacher behavior has been investigated using numerous observation instruments. When measures of teacher behavior are examined over more than one point in time, they have not been particularly stable. It may be that certain characteristics of the teaching situation affect the stability of particular teaching acts over time. For example, if a teacher is observed teaching high ability students at two different points in time, the frequency of, say, higher-order questions may be fairly stable over occasions. However, if this teacher is observed teaching high ability students at one point in time and low ability students at another point in time, the frequency of higher-order questions for that teacher may be very unstable over observation occasions.

Rosenshine and Furst (1973) defined the stability problem as one of "representativeness" or determining whether a sample of observed classroom transactions is a trustworthy representative sample of total behavior. Armstrong, DeVault, and Larson (1967) attempted to address this problem by determining the amount of observation time necessary to adequately sample a teacher's verbal communication using the Wisconsin Teacher Education Research Project (TERP) observation system and the Flanders system. They concluded that when sampling a teacher's behavior during reading instruction one should observe for approximately 150 minutes in order to obtain an adequate sample of teaching behavior in a one-week period. If the observations occur during arithmetic or social studies instruction, they concluded that the teacher's behavior should be sampled for approximately 120 minutes. Unfortunately, due to problems of sampling and design, the results and conclusions from this study are open to

several possible interpretations. However, this study does illustrate an important problem for future research on teaching.

Data on the stability of measures of teacher behavior, reviewed below, are examined to determine the extent to which they generalize over different teaching situations. As with the studies of teacher effectiveness, situational and observational facets have not been systematically varied in the literature to test the effect of a particular facet on stability. Therefore, conclusions about stability for a particular facet are confounded by unsystematic variation of other facets. However, results can be examined for patterns which suggest consistency or inconsistency.

Studies which present data on the stability of measures of teacher behavior are summarized in Table 3. The same assignment rules as described for Table 1 were used. The only facet which differs in Table 3 is the type of measure facet. This facet refers to whether the measure of teacher behavior is a frequency count of specific, denotable behaviors such as frequency of verbal reinforcements, a factor score representing a cluster of teacher behaviors, or a more general rating of such variables as clarity or enthusiasm which require the observer to process many cues before making a decision. The primary measure of stability of teacher behavior in these studies consisted of correlations between measures of teacher behavior obtained at two or more points in time.

- -

Insert Table 3 about here

- -

In the Hiller, Fisher, and Kaess (1969) study, videotapes from the Belgard, et al. (1968, see Table 2) study were examined for significant

verbal characteristics of the experienced teachers' lecture performance. The teachers delivered two different 15-minute lessons about the economic, social, and political conditions of Yugoslavia and Thailand to their twelfth-grade classes on successive days. Hiller, et al. used a computer program to identify the frequency of 35 categories of verbal responses (words, phrases) found in the Hiller Explanation Dictionary. The 35 categories dealing roughly with teachers' explanations were consolidated into a set of five factors based on the experimenters' a priori theoretical expectations for an optimal grouping of the categories. Correlations of these factors across observation occasions were reported. The two factors exhibiting greatest stability were verbal fluency (0.76) and vagueness (0.84). The other factors showed only moderate stability: optimal information amount (0.37), knowledge structure cues (0.29), and interest (0.20).

The high stability coefficients for verbal fluency and vagueness are difficult to interpret. The authors noted that vagueness arises as a speaker commits himself to deliver information he can't remember or never really knew. Therefore, it may be that the high stability of vagueness in this study is accounted for by the fact that the subject matter of these lessons was unfamiliar to the teachers. The authors also suggested that the verbal fluency factor may be a correlate of vagueness. That is, low verbal fluency, marked by hesitations in speech, may be correlated with vagueness. Therefore, the high stability of this factor may also be due to the unfamiliarity of the content being taught. However, this explanation is not confirmed by the data. The correlation between fluency and vagueness was -0.08 in the Yugoslavia lesson and -0.24 in the Thailand lesson.

In the Emmer and Peck (1973) study, videotapes of 28 fifth- and eighth-grade teachers were taken on five occasions. Each fifth-grade class was observed during mathematics, social studies, and science lessons. Eighth-grade classes included approximately equal numbers of mathematics, social studies, English, and science classes. Five systems were used to code the observations: Observation Schedule and Record (OScAR 5), Fuller Affective Interaction Records (FAIR), Cognitive Components System (CCS), Brophy-Good Dyadic Interaction System, and the Coping Analysis Schedule for Educational Settings (CASES). (Only the first four observation systems are considered here since the fifth system dealt exclusively with student behavior.) Data from each of the different systems were factor analyzed. (Eight to ten factors were identified for each of the four observation systems.) Intraclass correlations over one and five observation periods were computed for each factor in each observation system.

In order to examine the pattern of stability, we classified factors from each system into one of the following categories of variables: cognitive, affective, style, classroom management, and interpersonal. These categories of factors and their corresponding stability coefficients are presented in Table 3. The first category, cognitive variables, includes factors heavily loaded on behaviors relating to academic activities and the content being taught. For example, the problem-solving factor reflects teacher behaviors which initiate and structure problems for the class to consider. The stability coefficients for all but one of these cognitive dimensions are very low, ranging from below 0.10 to 0.18. (Emmer and Peck did not report intraclass correlations below 0.10 for factors in their table; hence, we say below or <0.10.)

The problem-solving factor showed slightly more stability (0.27).

The affective category refers to factors which are heavily loaded on the teacher's expression of feelings to students and teacher responsiveness to student feelings. For example, the positive affect dimension includes the teacher's praise or encouragement of pupils. The stability coefficients for these factors are also generally low (0.11 to 0.19).

The style category refers to factors which reflect a preferred orientation or mode of instruction on the part of the teacher. Three styles were identified. The first set of four factors reflected an orientation to student ideas rather than teacher ideas. The stability coefficients for these factors ranged from <0.10 to 0.22. A second set of factors was concerned with teachers' use of convergent questions versus divergent questions. They seemed to reflect an underlying dimension representing a focus upon single, correct answers rather than divergent discussion. The stability coefficients for these two factors were 0.36 and 0.14. The last set of style factors reflected teacher lecturing and indicated activities in which the teachers' presentation of content is the overriding concern. These factors showed the highest stability of all factors (0.28 to 0.51).

The next category of variables relate to classroom management. Two factors, procedural interaction and controlling, are heavily loaded on teacher behaviors which are procedural and nonsubstantive in nature. These factors show low stability (0.11 to 0.13). The other two factors in this category, desist statements and criticizing and asking for a behavior change, involve teacher requests for students to stop doing something. These factors were somewhat more stable (0.34 to 0.39).

The last category of factors concerns interpersonal variables

relating to specific kinds of teacher-pupil interactions. The stability for these factors was also rather low (<0.10 to 0.29).

In general, the factors isolated by Emmer and Peck range from unstable to moderately stable. The set of factors appearing to be most stable were those relating to teacher lecture. As Emmer and Peck pointed out, the class conditions in which observations were made were not homogeneous over time. Many facets, such as grade level, type of content, and similarity of content, varied from observation occasion to observation occasion. For example, in the fifth-grade classes, type of subject matter (e.g., mathematics, science) varied from occasion to occasion while in the eighth-grade classes the type of subject matter remained the same with individual topics differing. Furthermore, data from both grade levels were combined in the factor analyses. It may be that the unsystematic variation of these facets over observation occasions accounts for the low stability observed and that teacher lecture is less affected by the variation of these situational facets than the other behavior observed.

In both the Hiller et al. and Emmer et al. studies, frequency count variables were clustered either conceptually (Hiller et al.) or empirically (Emmer et al.) Questions about the consistency and replicability of the clusters aside, problems of interpretation arise. Unlike high inference variables, such as ratings of warmth where specific definitions for rating exist, clear definitions of clusters of variables do not exist. How are the variables within a cluster to be put together by the teacher into smooth teaching performance? How should a teacher be trained to perform the teaching represented by clusters? How should he or she be evaluated? While parsimony is welcome,

especially with respect to observation systems of teaching, the techniques used in these two studies may achieve this goal at too great a cost. The teaching process which was the original focus of the studies may be masked by conceptual or statistical factor analysis. Statistical techniques from information theory or from statistical theories of runs or stochastic processes might be better suited to identifying patterns of teaching acts without compromising the process itself (cf. Shavelson, 1973).

Moon (1969, 1971) observed the verbal behavior of 32 primary-grade teachers during science lessons. Half of the teachers had participated in Science Curriculum Improvement Study (SCIS) workshop which emphasized manipulation of materials and child-to-child communcation as an integral aspect of science lessons. The teachers using SCIS materials were observed on five occasions while the teachers teaching conventional science were observed on two occasions. The Flanders' instrument was used to collect interaction data and stability coefficients were calculated for two interaction ratios and five categories of questions for each group of teachers (see Table 3).

The I/D ratio is defined as the amount of indirect teacher influence in verbal classroom behavior divided by the amount of direct teacher influence. Teacher talk refers to the percentage of time teachers spent talking during science activities. The stability coefficients were fairly high for the teachers teaching conventional science (0.64 for I/D ratio and 0.76 for teacher talk) but very low for the teachers teaching SCIS science (0.18 for I/D ratio and 0.12 for teacher talk).

The following types of questions, categorized by Moon as lower-order showed virtually no stability with the SCIS teachers: questions which ask students to recall facts (0.12) and questions which ask students

to see relationships (0.05). These questions showed moderate stability with the conventional science teachers: 0.42 for recall questions and 0.43 for relationship questions. Two types of questions, categorized by Moon as higher-order, also showed essentially zero stability for both groups: questions which require observations (0.05 for SCIS teachers and 0.17 for conventional teachers) and questions which require testing hypotheses (0.00 for both samples). One type of higher-order question, questions which require generating hypotheses, showed somewhat more stability with both samples (0.19 for the SCIS sample and 0.17 for the conventional sample). It should be noted that questions calling for seeing relationships, hypothesizing and testing hypotheses are extremely infrequent events. Means for the frequency of occurrence of each of these question types ranged from 0.00 to 1.40 on any given occasion. The low correlations for these behaviors may reflect this restriction of range. The higher stability of the control teachers on some behaviors may reflect the fact that these teachers were teaching a familiar curriculum in their customary way, while the SCIS teachers were teaching an unfamiliar curriculum using somewhat new techniques for the first time.

Campbell (1972) observed five junior high school science teachers near the end of the spring semester and midway through the fall semester of the following academic year. Only teachers teaching students in the low achievement tract were studied. Three interaction ratios were calculated for each teacher on both observation occasions from data obtained with Flanders' system. From these data we calculated Spearman rank-order correlations for each interaction ratio.

Two interaction ratios involving indirect and direct teacher behavior were calculated by Campbell. The I/D ratio included the following

indirect (I) teaching acts: teacher accepts student feelings and student ideas, teacher praises, and teacher questions. Direct (D) teaching acts were: teacher gives directions, teacher criticizes, and teacher lectures or gives own opinion. The stability coefficient for this ratio (I/D) was 0.10. The revised i/d ratio did not include teacher questioning or lecturing behavior and is considered to be more sensitive to the affective areas of a teacher's behavior. The stability coefficient for this ratio i/d) was -0.90!

The third interaction ratio was the ratio of student talk to teacher talk (S/T). This ratio (S/T) showed the greatest stability over time (0.80). In all cases, these ratios were extremely low, indicating that the teachers dominated the verbal behavior of their classes. The high stability coefficient indicates that this pattern was consistent over time. This finding is subtantiated by Bellack et al. (1966) who found that 80% of all classroom time is spent in teacher talk.

The pattern of stability coefficients for the interaction ratios of indirect to direct teacher behavior is difficult to interpret. The first observation of teachers occurred during the final teaching weeks of the spring semester, and the second observation was recorded midway through the fall of the following academic year. As Campbell pointed out, the spring observations may reflect the cumulative fatigue of nine months of teaching, while the fall observations show a more normal pattern of interaction. Therefore, the instability of the I/D ratio and the i/d ratio may be due, at least in part, to the timing of the observations.

In the Friedman (1973) study, 18 teachers of grades one, three, five, or seven were observed on four separate occasions over a six-month

period. The Observation Schedule and Record (OScAR) was used to record pupil-initiated interchanges and teacher reinforcement of spontaneous student behavior. Only those interactions pertaining directly to classroom content (substantive interchanges) were recorded. The "exit" from an interchange occurred when the teacher either evaluated the child's verbalization or continued without evaluating it. Observation records were screened for six types of teacher exits to an interchange. Stability coefficients were calculated for each of the six exits and are reported in Table 3. The "acknowledged" exit (the teacher responds to what the child has said but does not indicate whether it is correct or not) and the "not evaluated" exit (the teacher makes no response to a child's verbalization) are the most stable (0.61, 0.58). The "approved" exit (the teacher indicates that what the child has said is acceptable but no enthusiasm or praise is given) and the "neutrally rejected" exit (the teacher indicates that what the child has said is unacceptable but no disapproval is expressed) are moderately stable (0.49, 0.48). The least stable exits were the "supported" exit (the teacher enthusiastically indicates that what the child has said is correct or acceptable) and the "criticized" exit (the teacher clearly indicates with disapproval that what the child has said is incorrect or unacceptable) (0.37, 0.26).

In order to analyze variations in teacher reinforcement of interchanges, a set of five orthogonal reinforcement categories was derived from the six original OScAR exits. These statistically orthogonal categories were scored as combinations of the original six exits. For example, the following weighting of the original six variables comprised the transformed variable, feedback: supported (1/2), approved (1/2), acknowledged (-1), not evaluated (-1), neutrally rejected (1/2), and

criticized (1/2). In general, the stability coefficients were higher for the transformed exits than the original exits (see Table 3). Encouragement 1 and encouragement 2 showed the highest stability (0.70, 0.73) with the feedback and positivity categories showing somewhat less stability (0.63, 0.61). The lowest stability was found for the enthusiasm exit (0.53).

Trinchero (1974a, b) examined the stability of teacher trainees' use of positive reinforcement, higher-order questions, and probing questions over a nine-month period. Observations occurred after technical skills training in August and at the end of student teaching in May. Teacher trainees were randomly assigned to teach a 40-minute English or social studies lesson to student volunteers from secondary schools in the Palo Alto, California, area. Students were randomly assigned to groups of approximately 22 each. The English lessons focused on some aspect of semantics, while the social studies topics dealt with economics. Each teacher was given an outline of the topic content, a list of topic objectives, and a copy of the topic test beforehand. Videotapes of these teaching sessions were rated for teacher acts reflecting use of reinforcement to increase student participatory behavior (positive reinforcement), teachers' use of questions which call for inferential responses (higher-order questions) and teachers' use of questions which require the student to go beyond his first response (probing). Stability coefficients were calculated for these teaching acts.

The stability coefficient for total reinforcement, the summation of frequency scores for all reinforcement categories, was 0.64. However, since there was a significant difference between the means of the English and social studies teachers on this variable, we calculated Spearman rank-order correlations separately for each subsample. These coefficients

were 0.47 for the English teachers and 0.63 for the social studies teachers.

Stability coefficients were also calculated with frequency scores and proportions for each of five categories of reinforcement (see Table 3). Positive nonverbal reinforcement was most stable (0.74). Separate calculations for each subsample yielded coefficients of 0.57 for English teachers and 0.64 for social studies teachers. Moderate stability coefficients were obtained for positive verbal reinforcement (0.44) and qualified verbal reinforcement, i.e., statements containing both positive and negative elements (0.48). However, when stability coefficients are calculated separately for positive verbal reinforcement for each subsample of teachers, an extremely low coefficient was found for English teachers (0.04), while a much higher one was found for the social studies teachers (0.57). The least stability was found for negative verbal reinforcement (0.13) and negative nonverbal reinforcement (-0.15). Stability coefficients for proportion scores were similar to coefficients for frequency scores (see Table 3).

Stability coefficients were also reported for frequency scores and proportions of higher-order questions (see Table 3). Very low stability (0.16 to 0.17) was observed for frequency of teacher-initiated questions, lower-order questions, and higher-order questions. Coefficients for proportions scores were much higher but in opposite directions (-0.52 for proportion of lower-order questions and +0.54 for proportion of higher-order questions). No explanation of these data are given by Trinchero; we are just as perplexed by them.

Stability coefficients for frequencies and proportions of probing are also shown in Table 3. The frequency of pupil-initiated talk seems

to be much less stable (0.04) than the frequency of opportunities to probe (0.24). The frequency of total probes and individual probes are moderately stable (0.40, 0.31), although the proportion scores for these variables are somewhat less stable (0.35 for total probes, 0.20 for individual probes).

Three specific types of probing are reasonably stable over time: individual redirects (0.54), group redirects (0.43), and clarifying and prompting questions (0.62). However, virtually no stability over time was shown for questions involving repeating, rephrasing, and refocusing (-0.09) or critical awareness (-0.05). Similarly, stability coefficients for proportion scores for all five types of probing questions were low (-0.29 to 0.14).

Borg (1972) observed 38 teachers immediately after completion of a minicourse designed to develop discussion skills and again four months later. Each teacher was instructed to prepare a 20-minute class discussion lesson based on content from current class work. Lessons were videotaped and later analyzed to determine the teacher's use of skills taught in the minicourse. We computed stability coefficients for these behaviors (see Table 3).

Two behaviors relating to general questioning showed low to moderate stability: teacher questions (0.14) and average length of pause (0.27). Two measures of questioning, "words in pupil response to information level questions", which was interpreted as indicating teacher's asking for sets of related facts and proportion of higher-order to fact questions showed virtually no stability (0.01; 0.07). Three types of probing questions showed moderate stability: redirection (0.40), prompting (0.30), and clarification (0.55).

Three skills considered undesirable in discussion showed low to moderately high stability: repeating own question (0.23), repeating pupil answer (0.22), and answering own question (0.10). Percentage of teacher talk as measured by the Flanders' system was also moderately stable (0.38).

In the Ehman (1972) study, 84 preservice secondary teachers (undergraduates) taught seven microteaching lessons to groups of four to six ninth-grade students as part of a social studies methods course. The first lesson, taught only to peers, was aimed at familiarizing students with the methods and mechanics of microteaching and was not included in the present analysis. The remaining six lessons were taught in pairs at about three-week intervals through the semester. In the first pair of lessons, the preservice teacher invented a lesson in which one or more historical or social science concepts were taught. Two days later, the same lesson was retaught under similar conditions to a new group of students. The second pair of lessons involved eliciting student hypotheses in response to a problem while the third set concerned the analysis of a value proposition or controversial issue. Audiotapes of each lesson were coded using a modified version of the Flanders' system, and stability coefficients were computed for each set of lessons (see Table 3).

The praises encourages category showed moderate stability (0.41 to 0.47), while the accepts uses student ideas category was less stable (0.08 to 0.47). Two questioning categories were fairly stable: overall questioning (0.33 to 0.52) and cognitive-memory questions (0.47 to 0.52) with one exception (-0.01). The other question types showed less stability (-0.01 to 0.35). Closed teacher questions or the ratio of cognitive-memory and convergent questions to total questions also showed low

stability (0.07 to 0.26).

The lecturing category was fairly stable (0.58 to 0.60) while the gives directions category was less consistent (0.15 to 0.44). Stability coefficients for teacher criticism were low (-0.04 to 0.25). Productive silence or confusion was moderately stable (0.49 to 0.55), while unproductive silence or confusion was much less consistent (0.14 to 0.36). Stability coefficients for the I/D ratio were quite high (0.57 to 0.66).

Ehman also computed stability coefficients between adjacent occasions for data from the Bellack et al. (1966) study in which 15 high school teachers were observed on four occasions (see Table 3). Teachers were given a pamphlet and teachers' guide on international economic problems and were asked to base their four lessons on this material. Typescripts were prepared from audiotapes of these lessons which were then analyzed using the Bellack system.

The ratio of teacher talk to student talk showed moderate to high stability over the four experimental sessions (0.41 to 0.84). Structuring moves by the teacher which set the context for subsequent behavior were also highly stable (0.72 to 0.76). Soliciting moves which include questions, commands, and requests showed high stability in two instances (0.61, 0.74) but not in the third (-0.09). Coefficients for teacher responding which involves a response to a soliciting move and teacher reacting which serves to modify or evaluate previous statements were less consistent (0.22 to 0.66).

Marzano (1973) randomly selected 27 teachers from a group of 50 first-grade teachers attending a two-day orientation workshop in which the teachers were trained to use the Distar Language 1 program. The Distar programs are designed to teach basic concepts at a fast pace so

that children who start out behind average youngsters have the opportunity to catch up. The program is unique in that the teacher's behavior is programmed. The teacher's role is one of teaching a particular format to mastery by saying and doing exactly what the materials prescribe.

Each teacher was audiotaped while teaching one lesson during each of five one-week segments. The five taping segments were equally spaced throughout a six-month period (October to March) so that lessons taught from Books A, B, and C of the Distar Language 1 program were observed. The variable observed was the degree to which the teacher followed the Distar format in group activities and individual activities.

Analysis of variance procedures (teachers x occasions x books) were carried out for each type of activity. Unbiased estimators of each of the components of variance were calculated and then used to construct estimates of the generalizability coefficients given in Table 3. The values of this coefficient were 0.52 and 0.54 for group and individual activities respectively. Even though, in computing these coefficients, Marzano omitted the variance component due to the occasion by book interaction in his definition of observed score variance, this did not affect the stability estimate since the interaction variance equalled zero. When teacher scores are averaged over the three books and five occasions, the coefficients are 0.92 for group activities and 0.92 for individual activities.

Wallen (1970) observed 33 teachers teaching "academic" lessons on ten occasions using the STEP Observation Schedule which includes ratings taken after every five minutes of observation and global ratings taken after the entire observation period (20-35 minutes). Five observations were obtained in the fall and five in the spring. Mean scores across

each set of five observations were obtained for each teacher and correlation coefficients were calculated between the fall and spring mean scores (see Table 3).

The stability coefficients for the global ratings were very high, ranging from 0.71 to 0.92 with one exception (0.03). In general, the five-minute ratings were much less stable. For example, teacher methods for focusing pupil's attention on the learning activity and allowing them to respond overtly were generally unstable (-0.02 to 0.21) with three exceptions (0.38, 0.42, 0.67). Clear instances of teacher individualization based on diagnosis of student needs were stable (0.62, 0.87) while other individualization variables were not (-0.09 to 0.21). Two feedback variables, feedback from teacher and no provision, were moderately stable (0.34 to 0.43) while the others were not (-0.22 to 0.00).

Direct controlling responses or ignoring of irrelevant behavior were also stable (0.61 to 0.77) while direct controlling responses to disruptive behavior and indirect controlling responses to both irrelevant and disruptive behavior were less stable (-0.13 to 0.42).

Two studies (Brophy and Evertson, 1974a, b; Sandoval, 1974) were located which are potential sources of data on stability of teacher behavior; however, no conclusions can be drawn at this point. Brophy and Evertson studied the teaching behavior of second- and third-grade teachers from both high SES and low SES schools selected because they had previously shown relatively high consistency in the amount of student learning gain that they produced on standardized achievement tests. The teachers were observed on 18 occasions over a two-year period using a variety of high-inference and low-inference measures. Pearson correlations between these process measures and student outcome measures as well as

multiple regression analyses designed to indicate nonlinear relationships between these measures were done. Unfortunately, correlations were not reported for the measures of teacher behavior over observation occasions. However, with the variety of teacher behavior measures used and the large number of observation occasions, this project is a rich source of data for examining the stability of teacher behavior measures over time. Furthermore, the effect of two systematically varying facets, type of measure (high-inference or low-inference) and time of day (morning, afternoon, or reading group), could be examined with four populations of teachers (second and third grade; Title 1 and non-Title 1 schools).

In the Sandoval (1974) study, 12 teachers were videotaped during both reading and mathematics on four occasions during the school year. Lessons taught on the second, third, and fourth observation occasions were coded using the Teacher Instruction Lexicon, Teacher Feedback Lexicon, and the Observation Classroom Summary Record from the APPLE system and a videotape rating of teacher behaviors and characteristics. Lessons from the second and third observation occasions were also coded using interaction analysis. Lessons from occasions one through three were coded with a videotape rating of teaching style as well. (Data from the first observation occasion was discarded for use with the other observation systems because of the poor quality of the videotape.)

The data from these observations were analyzed with an analysis of variance nested design (teachers nested within observers). However, this design does not permit an examination of the stability of teacher behaviors over time. The main effect for teachers and the interaction of teachers and observation occasions is confounded by the effect of raters.

Summary: Studies of Teacher Behavior

In order to determine the effects of various situational and observational facets on the generalizability of measures of teacher behavior, teaching acts which were defined similarly were examined as a group (see Table 4). If two or more variables from one study fell into a cluster, correlational evidence, if available, was examined to verify the clustering. For example the cluster, teacher presentation, contained three variables (teaching acts) from the Emmer and Peck (EP) study; these three variables were highly correlated with each other. In addition, the cluster included lecturing (Wallen; Ehman), teacher talk (Moon; Borg), student/teacher talk ratio (Campbell; Bellack), and three I/D ratios (M, C, E).

- -

Insert Table 4 about here

- -

In clustering these teaching acts, an already apparent problem became critical. It was impossible to verify that measures of the same variable, observed by two different investigators using the same instrument on questionably similar populations of teachers, were actually measuring the same thing. More commonly, the same or slightly different variable was measured by two different instruments in two different studies. The validity of interpretations of such measurements is suspect, at best. This is an issue of the validity of construct interpretations of these measures. In addition, standardization[2] of measurements, or the lack thereof, has played a major part in inhibiting the advance of knowledge about teaching. The lack of evidence on construct validity and standardization makes the validity of our interpretations of teacher variables and so our clustering quite tentative.

In drawing together evidence on the generalizability of various measures of teacher behavior, a distinction between global ratings and other types of measures will be maintained. Only Wallen's (1970) study reported the stability of global ratings. The stability of these ratings stands in striking contrast to the stability of most other measures, as will be seen. Whether the stability is in the teachers observed or is in the beholder's (observer's) eye, cannot be untangled in this review.

The frequency count, factor score, and short-term rating measures were clustered into 12 categories: teacher presentation, presentation of content, teacher questions, teacher feedback, teacher motivational skills, expressive teaching style, student-centered teaching style, interpersonal behavior, affective style, teacher following procedures, teacher's classroom management, and teacher individualization.

Teacher Presentation

Twelve variables were grouped into the teacher presentation cluster (see Table 4). In most cases, measures of teacher presentation are reasonably stable. Since the SCIS teachers in the Moon study received training which encouraged student inquiry rather than teacher lecture, instability in measures of their presentation, compared to conventional teaching, might be expected. The low coefficients from Wallen and Campbell, then, need some explanation--possibly in terms of differences in facets from the other studies. Students in the Wallen study were from the inner-city and students in the Campbell study were low achievers. Information about students in the other studies is not given. One might speculate that the type of student population in the Wallen and Campbell studies did not allow stable lecturing behavior. That is, lecturing may

be difficult or inappropriate with low achieving or poorly motivated students. However, this represents conjecture, at best. Furthermore, the Campbell coefficient is based on only five teachers.

Presentation of Content

The second cluster, presentation of content, might be considered a further delineation of the teacher presentation cluster--various types of presentations fall into this cluster. Of the 13 variables in the cluster, three relating to explaining in general, four related to higher-order cognitive processing (as opposed to factual recall), one variable related to lower-order cognitive processing, and one to structuring of the lesson. Finally, four variables related to the teacher's speaking style (cf. Table 4). While teacher presentation seems to be somewhat stable, the stability coefficients of variables further delineating teacher presentation--presentation of content--appear to be quite low, with the exception of the structuring and speaking style variables. The low stability coefficients in this cluster came predominantly from the Emmer and Peck study. The major difference in facets between this study and the other three studies (H, E, Be) was the type of subject matter (EP=mathematics, social studies, English, science; H, E, Be=social studies). Perhaps the subject matter in the Hiller et al., Ehman, and Bellack studies permitted consistency of presentation and speaking style that would not have been found with Emmer and Peck's different subject matters all combined. Or perhaps the difference lies in the qualitatively different variables examined by Emmer and Peck in contrast to the structuring and speaking style variables in the other studies.

Teacher Questions

The third cluster, teacher questions, contains 24 variables. Five related to frequency of questions, 11 referred to particular types of questions, and eight related to clusters representing patterns of questioning (see Table 4).

The data from these seven studies indicate that teacher questioning behavior is unstable. The exceptions for overall frequency of questions are: Ehman (r=0.33, 0.34, 0.52) and Bellack (r=-0.09, 0.61, 0.74). The exceptions for particular types of questions are Moon's conventional teachers asking what he termed lower-order questions (0.42, 0.43) and Bellack's teachers asking cognitive-memory questions (-0.01, 0.47, 0.52).

Examination of the facets of studies showing stability (E, Be) and instability (T, W, B) in overall questioning shows that stable questioning behavior resulted when the same or similar topics from a particular subject matter, social studies, were taught. Instability occurred when subject matters were combined and different topics taught (T=English, social studies; W="academic" lessons; Borg=unknown). A similar pattern occurred when lower-order questioning was examined. Stability in lower-order questions occurred when one subject matter (M=science; Be=social studies) with the same or similar topics was taught. Instability occurred when subject matters were combined and different topics taught (T=English, social studies; EP=mathematics, social studies, English, science; B=unknown).

A conjecture about differences in stability between Moon's SCIS and conventional teachers is also possible. Since Moon's SCIS teachers received training which encouraged asking higher-order questions, the

training may have interfered with their normal lower-order question asking. If training were less than successful, this might account for the instability in higher-order questions.

Teacher Feedback

Another cluster of variables deals with the teacher's use of feedback (reinforcement) in the classroom. Of the 23 variables relating to feedback, five referred to total reinforcement, seven dealt with positive feedback, four related to neutral feedback, and seven variables were related to criticism or negative feedback (see Table 4). A consistent pattern of stability coefficients emerges from the six studies examining teacher feedback. Teachers' use of positive and neutral feedback appears to be moderately stable. Teachers' use of criticism or negative feedback appears to be unstable. This pattern seems to be stable over variations in such facets as grade level, similarity of students, and differences in subject matter similarity, at least to the limited extent such a generalization is warranted.

Motivational Skills

The cluster, teacher motivational skills, contains four variables (see Table 4). It should be noted that selling is a very infrequent event. Therefore, the low stability coefficient for this variable may reflect this restriction of range. The variables fit quite loosely in the cluster, come from only two studies, and thus do not deserve further comment.

The five clusters just discussed--teacher presentation, presentation of content, questions, feedback, and motivation--might be thought of as

relating to quality of interactive instruction. The next four clusters--expressive teaching style, student-centered style, interpersonal behavior, and affective teaching style--reflect a personal style in teacher-student interactions.

Expressive Teaching Style

Five variables comprise the expressive teaching style cluster (see Table 4). Since the Friedman variables are orthogonal and the two Emmer and Peck variables share only a small portion of variance, these variables represent different aspects of expressive style. All variables represent combinations of low inference variables, and the pattern of coefficients reflect the general magnitude of coefficients reported in the two studies. Differences and similarities in facets do not seem to account for differences in the magnitudes.

Student-Centered Teaching Style

The cluster, student-centered teaching style, contains six variables. Five of them reflect a student-centered orientation, and the last variable represents the opposite of a student-centered orientation (see Table 4). Student-centered teaching style, based on data from three studies, appears to be relatively unstable with the exception of one coefficient in the Ehman study (0.47). This coefficient is not unexpected since it occurred in a lesson involving an analysis of values which easily lends itself to accepting and using student ideas. The last variable representing teacher control of interaction suggests that teachers might be stable in their control of interaction and somewhat unsystematic in allowing students to control the interaction. But this interpretation hinges on

one coefficient and represents speculation, at best.

Interpersonal Behavior

The cluster, interpersonal behavior, contains 26 variables (see Table 4). The first four are from EP, and the next seven relate to teacher provisions for overt pupil responses. Nine variables are from Trinchero's probing category. Two reflect classroom opportunities for probing, and the others reflect the teacher's response to a student response. Another three variables are from Borg and are similar to Trinchero's probes. The last three variables are also from Borg and represent teacher responses during discussion. While these variables fit loosely into this cluster, the variables from the EP and W studies seem unrelated to those from the T and B studies. This lack of overlap, along with a lack in consistency in the magnitude of coefficients, makes these findings difficult to interpret. The probing variables, however, in both the T and B studies appear to be moderately stable. Since these variables come from only two studies, no statement about the generalizability over facets can be made.

Affective Style

The last cluster in the style series--affective style--contains three variables (see Table 4). The coefficients from EP are quite low; the coefficient from C is based on only five teachers. Affective variables in these studies, then, appear to be unstable over time.

The last three clusters are administrative in nature (see Table 4).

Following Procedures

One cluster refers to teachers following procedures. The differences between the results of Marzano and Wallen are probably due to the differences in the nature of the variables. Marzano's teachers were given a Distar curriculum which prescribed each teacher task. The Wallen variables come from less controlled teacher-student interaction.

Classroom Management

The next cluster, classroom management, includes administrative variables and classroom control variables. Of the 19 variables in the cluster, two related to classroom administration, seven variables related to direct controlling responses to undesirable student behavior, and ten variables related to indirect controlling responses to undesirable student behavior.

The administrative variables were moderately stable in one study (E) but not in the other (EP). However, the data come from only two studies, and it is unclear whether this difference resulted from differences in facets or in the nature of the variables. In general, direct controlling responses of the teacher to undesirable student behavior appear to be moderately stable. Indirect controlling responses appear to be unstable with two exceptions: praising desired behavior (0.42) and ignoring (0.66) in response to irrelevant student behavior. However, the other indirect controlling behaviors were extremely infrequent. Therefore, the low coefficients for these behaviors may reflect this restriction of range.

Teacher Individualization

The last cluster, teacher individualization, relates to teacher provision for individual differences among students and includes seven variables, all from the Wallen study. The only variables in this cluster which appeared to be stable were instances of individualization and ability grouping which were rated as showing clear indication of teacher diagnosis of student needs. It is unclear whether this stability indicates consistency of teacher diagnosis or consistency of observer perceptions.

Global Ratings

Wallen (1970) was the only study to report stability of global ratings. Stability coefficients ranged from 0.71 to 0.92 for the variable, control of academic behavior (0.03). Instability in the latter rating might be expected since this variable related to the extent to which a teacher's lessons are highly structured and controlled and both subject matter and topics varied from observation to observation. The global ratings reported by Wallen are substantially more stable than the other types of measures in general and higher than those variables in clusters which consistently showed stable behavior (viz. teacher presentation, positive and neutral feedback). As noted above, we cannot determine whether this stability represents an attribute of the teachers observed or of the observer's perceptions.

In summary, global ratings appear to be the most stable of all measures examined. In addition, four clusters of variables appeared moderately stable over variations in facets: teacher presentation, positive and neutral feedback, probing, and direct teacher control (classroom management).

The stability of six clusters was unclear; that is, some variables within the cluster were stable while others were not. These clusters were: presentation of content, teacher motivational skills, expressive teaching style, teachers following procedures, classroom administration, and teacher individualization. The reasons for these differences in stability, however, is unclear. For example, any of the following explanations for these inconsistencies are possible: (1) lack of standardization of measures, (2) qualitative differences between variables within the cluster, or (3) variations in one or a combination of facets between studies.

The remaining clusters appeared unstable: teacher questioning, negative feedback, student-centered teaching style, interpersonal behavior, affective style, and indirect teacher control (classroom management). However, since facets were not systematically varied within studies, the limits of generalizability over particular facets are unclear.

Conclusions

A careful examination of the stability of measures of teacher effectiveness and teaching process suggests that generalizability may be extremely limited in an educational context. The lack of systematic variation in situational and observational facets and lack of standardization of measurements; however, makes our conclusions tentative, at best. With this in mind, the following conclusions are in order.

1. The long-term studies of teacher effectiveness show low to moderate stability coefficients; however, an explanation of the variation in these coefficients does not emerge when various facets of the measurements are examined. In order to explain what appears, at present, to be

unsystematic variation in stability coefficients, issues of internal and external validity of these studies must be addressed.

2. From the limited data available from the short-term studies of teacher effectiveness, the effects of teachers teaching the same content to similar students appears to be moderately stable and consistent over studies. The magnitude of stability coefficients for teachers teaching different content to the same or similar students varies greatly. When the effects of other facets are examined in an attempt to explain this inconsistency, the variability in stability coefficients appears to be independent of the variability in facets.

3. Of all the measures of teacher behavior examined, global ratings appear to be the most stable. It is unclear whether this consistency represents an attribute of the teachers observed or of the observers' perceptions. Frequency counts tended to have the lowest bound stability coefficients. In some cases these low coefficients could be explained by restricted range due to the low frequency of occurrence of certain teacher behavior (e.g., variables in the indirect teacher control cluster).

4. Four clusters of teacher behavior variables based on other measures appeared moderately stable over variations in facets: teacher presentation, positive and neutral feedback, probing, and direct teacher control (classroom management).

5. The stability of the variables in six of the clusters was unclear, that is, some variables were stable and some were unstable within the cluster: presentation of content, teacher motivational skills, expressive teaching style, teachers following procedures, classroom administration, and teacher individualization. Any of the following

explanations for this inconsistency are plausible: (a) lack of standardization of measures, (b) qualitative differences between variables within the cluster, or (c) variations in one or a combination of facets between studies.

6. Variables in six of the clusters appeared unstable: teacher questioning, negative feedback, student-centered teaching style, interpersonal behavior, affective style, and indirect teacher control (classroom management). Systematic variation of facets is required in order to explain this apparent instability.

FOOTNOTES

1 The term "facets" is used here to describe the conditions of observation; it subsumes such terms as "settings", "situations", (Rosenshine, 1970) and "contexts", (Rosenshine, personal communication). This term is borrowed from generalizability theory (Cronbach et al., 1972) which as guided this review. "Facet" is a general term referring to conditions of an observation of a certain kind. Thus, observations may be classified with respect to the facet of type of student, the facet of type of subject matter, the facet of observers, etc.

2 "A standardized test is one in which the procedure, apparatus, and scoring have been fixed so that precisely the same testing procedures can be followed at different times and places" (Cronbach, 1970, p. 27).

REFERENCES

Anastasi, A. Psychological Testing. New York: McGraw-Hill, 1968.

Armstrong, J. R., DeVault, M. V., & Larson, E. Consistency of teacher communication: a sampling problem in interaction analysis. Psychology in the Schools, 1967, 4, 1, 74-76.

Belgard, M., Rosenshine, B., Gage, N. L. The teacher's effectiveness in explaining: evidence on its generality and correlation with pupils' ratings and attention scores, in Gage, N. L., Explorations of the Teacher's Effectiveness in Explaining. Stanford: Stanford Center for Research and Development in Teaching, Stanford University, Technical Report No. 4, 1968.

Bellack, A. A., Kliebard, H. M., Hyman, R. T. and Smith, F. L. The Language of the Classroom. New York: Teachers College Press, 1966.

Berliner, D. C. and Ward, B. A. Proposal for Phase III Beginning Teacher Evaluation Study, Far West Laboratory for Educational Research and Development, 1974.

Borg, W. R. The minicourse as a vehicle for changing teacher behavior: A three-year follow-up. Journal of Educational Psychology, 1972, 63, 572-579.

Bracht, G. H. and Glass, G. V. The external validity of experiments. American Educational Research Journal, 1968, 5, 437-474.

Brophy, J. E. Stability of teacher effectiveness. American Educational Research Journal, 1973, 10, 3, 245-252.

Brophy, J. E., & Evertson, C. M. Process-product correlations in the Texas teacher effectiveness study: final report. Report No. 74-4, University of Texas at Austin, 1974.

Brophy, J. E., & Evertson, C. M. The Texas teacher effectiveness project: presentation of non-linear relations and summary discussion. Report No. 74-6, University of Texas at Austin, 1974.

Campbell, J. R. A longitudinal study in the stability of teachers' verbal behavior. Science Education, 1972, 56, 1, 89-96.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In Gage, N. L. (Ed.) Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Connor, A. Cross-validating two performance tests of instructional proficiency. UCLA. U.S.D.E. Project No. 8-I-174, 1969.

Creemers, B. Evaluation of Styles of Teaching Initial Reading: An educational investigation into the relationship between teachers' use of a specific method and pupil achievement (with summary in English). Utrecht: Drukkerij Elinkwijk B. V., 1974.

Creemers, B. and Weeda, P. Evaluation of teaching styles: An investigation into the relationship between pupil achievement and teachers' use of a method for teaching initial reading. Unpublished manuscript, State University of Utrecht, Netherlands, 1974.

Cronbach, L. J. Essentials of psychological testing. Third Edition, New York: Harper and Row, 1970.

Emmer, E. T., & Peck, R. F. Dimensions of classroom behavior. Journal Educational Psychology, 1973, 64, 2, 223-240.

Flanders, N. A., & Simon, A. Teaching effectiveness: A review of research, 1960-66. In Ebel, R. L. (Ed.) Encyclopedia of Educational Research. Chicago: Rand McNally, 1970.

Fortune, J. C. The generality of presenting behaviors in teaching preschool children. Unpublished report. Central Midwestern Regional Education Laboratory, St. Ann, Missouri. August, 1966, pp. 75-76.

Fortune, J. C. A study of the generality of presenting behavior in teaching. U. S. Office of Education Project No. 6-84 64, Memphis State University, 1967 (ERIC: ED 016 285).

Fortune, J. C., Gage, N. L., & Shutes, R. E. The generality of the ability to explain. Paper presented to American Educational Research Association, 1966.

Friedman, P. Relationship of teacher reinforcement to spontaneous student verbalization within the classroom. Journal of Educational Psychology, 1973, 65, 1, 59-64.

Gage, N. L. Paradigms for research on teaching. In Gage, N. L. (Ed.) Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Gage, N. L. Teacher effectiveness and teacher education: The search for a scientific basis. Palo Alto, California: Pacific Books, 1972.

Glass, G. Statistical and measurement problems in implementing the Stull Act. Conference on the Stull Act, Stanford University, 1972. (a)

Glass, G. The wisdom of scientific inquiry on education. Journal of Research in Science Teaching, 1972, 9, 3-18. (b)

Harris, A. J., Morrison, C., Serwer, B. L., & Gold, L. A continuation of the CRAFT project: Comparing reading approaches with disadvantaged urban Negro children in primary grades. Cooperative Research Project No. 5-0570-2-12-1. Division of Teacher Education of City University of New York, 1968. (ERIC: ED 020 297).

Heath, R. W. and Nielson, M. A. The myth of performance-based teacher education. Paper presented to American Education Research Association, 1973.

Hiller, J. H., Fisher, G. A., & Kaess, W. A computer investigation of verbal characteristics of effective classroom lecturing. American Educational Research Journal, 1969, 6, 661-675.

Justiz, T. B. A reliable measure of teacher effectiveness. Educational Leadership, 1969, 3, 49-55.

Marzano, W. A. Determining the reliability of the Distar instructional system observation instrument. Unpublished thesis, University of Illinois, 1973.

McNeil, J. D., & Popham, W. J. The assessment of teacher competence. In Travers, R. (Ed.) Second Handbook of Research on Teaching. Chicago: Rand McNally, 1973.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In Gage, N. L. (Ed.) Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

Millman, J. Psychometric characteristics of performance tests of teaching effectiveness. Paper presented to American Educational Research Association, 1973.

Moon, T. C. A study of verbal behavior patterns in primary grade classrooms during science activities. Unpublished doctoral dissertation, Michigan State University, 1969.

Moon, T. C. A study of verbal behavior patterns in primary grade classrooms during science activities. Journal of Research in Science Teaching, 1971, 8, 171-177.

Rosenshine, B. The stability of teacher effects upon student achievement. Review of Educational Research, 1970, 40, 647-662.

Rosenshine, B. and Furst, N. Research on teacher performance criteria. In Smith, B. O. (Ed.) Research in Teacher Education. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1971.

Rosenshine, B. and Furst, N. The use of direct observation to study teaching. In Travers, R. (Ed.) Second Handbook of Research on Teaching. Chicago: Rand McNally, 1973.

Sandoval, J. Beginning teacher evaluation study completion report, Task K: Sub-study of consistency of teacher behavior. Princeton, New Jersey and Berkeley, California: Educational Testing Service, 1974.

Scriven, M. The evaluation of teachers and teaching. California Journal of Educational Research, 1974, 25, 109-115.

Snow, R. E. Brunswikian approaches to research on teaching. American Educational Research Journal, 1968, 5, 475-489.

Snow, R. E. Representative and quasi-representative designs for research on teaching. Review of Educational Research, 1974, 44, 265-291.

Soar, R. S. An integrative approach to classroom learning. Public Health Service Grant No. 5-R11 MH 01096 and National Institute of Mental Health Grant No. 7-R11 MH 02045. Temple University, Philadelphia, Pennsylvania, 1966. (ERIC: ED 033 749).

Trinchero, R. L. The longitudinal measurement of teacher effectiveness. California Journal of Educational Research, 1974, 25, 121-127.

Trinchero, R. L. Three technical skills of teaching: Their stability and effect on pupil attitudes and achievement. Unpublished doctoral dissertation, Stanford University, 1974.

Veldman, D. J., & Brophy, J. E. Measuring teacher effects on pupil achievement. Journal of Educational Psychology, 1974, 66, 319-324.

TABLE 1
Long-Term Studies of Teacher Effectiveness

	STUDENTS				CONTENT		TEACHER	DEPENDENT VARIABLE				
Study	Assignment to Teacher	Grade Level	Similarity	Type of Population	Type	Similarity	N	Level of Experience	No. of Observation Occasions	Posttest Questions Provided	Type of Measure	Content	Stability Coefficient
Marsh et al. (1955)	non-random	military training	similar	air force recruits	airplane hydraulics	same	106	-	2	no	O-R	written test performance test combined test	0.34 0.32 0.38
Harris et al. (1968)	non-random	1st	similar	urban Blacks low SES	reading (new curriculum	similar	30	P	2	no	N-R	vocabulary word reading paragraph meaning spelling word study skills	-0.01 0.53 0.32 0.52 0.36
Harris et al. (1968)	non-random	2nd	similar	urban Blacks low SES	reading (new curriculum	similar	24	P	2	no	N-R	word knowledge word discrimination reading spelling	0.19 0.26 -0.08 0.21
Soar (1966)	non-random	3rd-6th	similar	upper-middle class high ability	reading, arithmetic	similar	55	P	2	no	N-R	achievement gain factor (vocabulary, reading, arithmetic, concepts, arithmetic problems)	0.09
Torrance & Parent (1966)	non-random	7th-12th	similar	-	mathematics	similar	63	P	2	no	N-R	mathematics	anecdotal report of little consistency
Beatty (1973)	non-random	2nd-3rd	similar	Title I or non-Title I	reading, arithmetic	similar	165	P (female)	3	no	N-R	median coefficients for word knowledge, word discrimination, reading, arithmetic	Grade 2: Title I 0.25, Non-Title I 0.42; Grade 3: Title I 0.39, Non-Title I 0.40

Long-Term Studies of Teacher Effectiveness
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	DEPENDENT VARIABLE: No. of Observation Occasions	DEPENDENT VARIABLE: Posttest Questions Provided	DEPENDENT VARIABLE: Type of Measure	DEPENDENT VARIABLE: Content	DEPENDENT VARIABLE: Stability Coefficients Grade 2 Title I	Grade 2 Non-Title I	Grade 3 Title I	Grade 3 Non-Title I
Veldman & Brophy (1974)	non-random	2nd-3rd	similar	Title I or non-Title I	reading, arithmetic	similar	115	P (female)	3	no	N-R					
												word knowledge	0.43	0.66	0.78	0.63
												word discrimination	0.36	0.74	0.26	0.49
												reading	0.24	0.66	0.50	0.23
												arithmetic computation	0.00	0.48	0.63	0.39
												arithmetic reasoning	-	0.61	0.27	0.64
												verbal average	0.35	0.71	0.65	0.38
												quan. average	-	0.59	0.50	0.75
												Total	0.19	0.69	0.54	0.65
Creemers (1974)		1st		middle class (Dutch)	initial reading		31		2		M-R O-R					

TABLE 2

Short-Term Studies of Teacher Effectiveness

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	DEPENDENT VARIABLE: No. of Observation Occasions	DEPENDENT VARIABLE: Posttest Questions Provided	DEPENDENT VARIABLE: Type of Measure	DEPENDENT VARIABLE: Content	DEPENDENT VARIABLE: Stability Coefficients
Fortune (1966a)	quasi-random	pre-school	three conditions: 1. similar	Headstart	social studies (transportation, community workers)	three conditions: 1. same	15 15	T	2	no	O-R	social studies	0.45 0.22
			2. same	Headstart	social studies (transportation, community workers)	2. different	15	T	2	no	O-R	social studies	-0.27 -0.08 0.12 0.40
			3. similar	Headstart	social studies (transportation, community workers)	3. different	15	T	2	no	O-R	social studies	0.21 -0.08 0.42 0.13 0.45 0.25
Fortune, Sage, Shutes (1966a)	random	7th,9th	1. similar		social studies	1. same	40	T	2	yes	O-R	social studies	0.51
			2. same		social studies	2. different	40	T	2	yes	O-R	social studies	0.03 0.13
			3. similar		social studies	3. different	40	T	2	yes	O-R	social studies	-0.10 0.01 0.43
Fortune, (1967)	quasi-random	4th,6th	1. similar	university-related	English, math, social studies	1. same	13 14 15	U	2	no	O-R	social studies English math	0.70 0.57 0.68
			2. same	university-related	English, math, social studies	1. different	13 14 15	U	2	no	O-R	social studies English math	-0.14 -0.43 -0.27 0.12 -0.17 0.04
			3. similar	university-related	English, math, social studies	1. different	13 14 15	U	2	no	O-R	social studies English math	-0.45 -0.42 0.42 -0.35 -0.07 0.53 -0.29 0.05 0.82

Short-Term Studies of Teacher Effectiveness
(Continued)

	STUDENTS				CONTENT		TEACHER	DEPENDENT VARIABLE				
Study	Assignment to Teacher	Grade Level	Similarity	Type of Population	Type	Similarity	N	Level of Experience	No. of Observation Occasions	Posttest Questions Provided	Type of Measure	Content	Stability Coefficients
Belgard, Rosenshine, Gage (1965b)	non-random	12th	same	-	social studies (economic, political, & social conditions of Yugoslavia & Thailand)	different	43	P	2	yes	O-R	social studies	0.47
Justiz (1969)	random	12th	1. same	-	varied (news story structure, computer punch cards)	1. different	7	T	2	no	O-R	varied	0.90[a]
			2. similar		varied (news story structure, computer punch cards)	2. different	10	T	2	no	O-R	varied	0.63[a]
Connor (1969)	random	senior high school	similar	-	varied (propegenda techniques, probability)	different	17	P (male)	2	no	O-R	varied	0.25[b] 0.35[a, b]
Millman (1973) Study 1	random	4th,5th 6th	same	urban, mostly minority	varied (light rays, adjectives, folkways: short lesson, folkways: long lesson)	different	34 34	T	2	yes	O-R	cognitive affective	0.02 0.34
			similar	urban, mostly minority	varied (light rays, adjectives, folkways: short lesson, folkways: long lesson)	different	34 34	T	2	yes	O-R	cognitive affective	-0.03 0.31

a) Data based on Glass' (1972) reanalysis of Connor's data with which we are in agreement.

b) Spearman rank-order correlation. Others are Pearson product-moment correlations.

Short-Term Studies of Teacher Effectiveness
(Continued)

	STUDENTS				CONTENT		TEACHER		DEPENDENT VARIABLE				
Study	Assignment to Teacher	Grade Level	Similarity	Type of Population	Type	Similarity	N	Level of Experience	No. of Observation Occasions	Posttest Questions Provided	Type of Measure	Content	Stability Coefficients
Study 2	random	2nd,3rd	similar	urban, mostly minority	varied (decoding, rhythms)	same	12 11	T	2	yes	O-R	cognitive affective	-0.23 0.26
			same	urban, mostly minority	varied (decoding, rhythms)	different	12 12	T	2	yes	O-R	cognitive affective	0.71 0.18
			similar	urban, mostly minority	varied (decoding, rhythms)	different	12 11	T	2	yes	O-R	cognitive affective	0.09 -0.22

TABLE 3
Studies of Teacher Behavior

	STUDENTS				CONTENT		TEACHER		DEPENDENT VARIABLE			
Study	Assignment to Teacher	Grade Level	Similarity	Type of Population	Type	Similarity	N	Level of Experience	No. of Observation Occasions	Type of Measure(s)	Behavior Observed	Stability Coefficients
Hiller et al. (1969)	non-random	12th	same	-	Social studies (economic, social, political conditions of Yugoslavia and Thailand)	different	23	P	2	factor score	verbal fluency	0.76
											optimal information amount	0.37
											knowledge structure cues	0.29
											interest	0.20
											vagueness	0.84

Studies of Teacher Behavior
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	DEPENDENT VARIABLE: No. of Observation Occasions	DEPENDENT VARIABLE: Type of Measure(s)	DEPENDENT VARIABLE: Behavior Observed	DEPENDENT VARIABLE: Stability Coefficients
Inner and Peck (1973)	non-random	5th,8th	same	-	math, social studies, science, English	different	28	P	5[a]	factor score	Cognitive:	
											problem-solving	0.27
											conceptual	<0.10
											descriptive vs. inferential	<0.10
											explanation	0.16
											association, drill	<0.10
											higher cognitive level student behavior	<0.10
											descriptive interchange	0.18
											Affective:	
											positive teacher affect	0.11
											teacher candor	0.19
											teacher support	0.15
											Style:	
											student ideas vs. teacher ideas	0.16
											student presents vs. teacher questions	0.16
											expansive teacher behavior	0.22
											student discussion	<0.10
											convergent vs. divergent teacher behavior	0.36
											convergent questions	0.14
											teacher informing vs. student statements	0.28
											teacher lecture	0.[illegible]
											teacher presents (i.e., lecture)	0.45

[a]One class observed only three times

Studies of Teacher Behavior
(Continued)

	STUDENTS				CONTENT		TEACHER		DEPENDENT VARIABLE			
Study	Assignment to Teacher	Grade Level	Similarity	Type of Population	Type	Similarity	N	Level of Experience	No. of Observation Occasions	Type of Measure(s)	Behavior Observed	Stability Coefficients
Emmer & Peck (1973) cont.	non-random	5th,8th	same	-	math, social studies, science, English	different	28	P	5[a]	factor score	Classroom Management:	
											procedural interaction	0.13
											desist statements	0.39
											controlling	0.11
											criticizing, asking for behavior change	0.34
											Interpersonal:	
											private pupil-teacher contacts	0.29
											teacher behavior when no pupil response	0.18
											teacher behavior when incorrect pupil response	0.18
											criticizing wrong answer	0.16
											teacher seeking correct answer	0.26
											criticizing pupil work	<0.10
											asking opinion question	0.14
											asking open question--correct answer given	0.23
											asking self reference question--teacher negates answer	0.13

[a]One class observed only three times

Studies of Teacher Behavior
(Continued)

	STUDENTS				CONTENT		TEACHER		DEPENDENT VARIABLE			
Study	Assignment to Teacher	Grade Level	Similarity	Type of Population	Type	Similarity	N	Level of Experience	No. of Observation Occasions	Type of Measure(s)	Behavior Observed	Stability Coefficients
Campbell (1972)	non-random	junior high	similar	low achievers	science	similar	5	P	2	frequency count	I/D ratio	0.10[b]
											i/d ratio	-0.90[b]
											S/T ratio	0.80[b]

[b]Spearman rank-order correlations. All others are Pearson product-moment correlations.

Studies of Teacher Behavior
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	No. of Observation Occasions	DEPENDENT VARIABLE: Type of Measure(s)	DEPENDENT VARIABLE: Behavior Observed	DEPENDENT VARIABLE: Stability Coefficients
Friedman (1973)	non-random	1st,3rd 5th,7th	same	-	-	-	72	P (female)	4	frequency count	Teacher Original Exits:	
											supported	0.37
											approved	0.49
											acknowledged	0.61
											not evaluated	0.58
											neutrally rejected	0.48
											criticized	0.26
										transformed scores	Transformed Exits:	
											feedback	0.63
											positivity	0.61
											enthusiasm	0.53
											encouragement 1	0.[illegible]
											encouragement 2	0.73

Studies of Teacher Behavior
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	DEPENDENT VARIABLE: No. of Observation Occasions	DEPENDENT VARIABLE: Type of Measure(s)	DEPENDENT VARIABLE: Behavior Observed	DEPENDENT VARIABLE: Stability Coefficients: SCIS Teachers	DEPENDENT VARIABLE: Stability Coefficients: Conventional Teachers
Moon (1969, 1971)	non-random	primary	same	-	science	similar	32	P	5(SCIS teachers 2(Control teachers	frequency count	ID ratio	0.18	0.64
											teacher talk	0.12[c]	0.76
											Teacher Questions:		
											fact or recall	-0.12[c]	0.42
											relationships	0.05[c]	0.43
											observations	0.05[c]	-0.17
											hypothesizing	0.19[c]	0.17
											testing hypothesis	0.00[c]	0.00

[c]based on Fisher's z-transformations

Studies of Teacher Behavior
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	DEPENDENT VARIABLE: No. of Observation Occasions	DEPENDENT VARIABLE: Type of Measure(s)	DEPENDENT VARIABLE: Behavior Observed	DEPENDENT VARIABLE: Stability Coefficients
Trinchero (1974)	random	9th,10th	similar	-	social studies English	different	20	T	2	frequency count	Reinforcement:	
											Total[d]	0.64
											English teachers	0.47[b]
											social studies teachers	0.63[b]
											positive verbal	0.44
											English teachers	0.04[b]
											social studies teachers	0.57[b]
											positive verbal/total verbal	0.53
											negative verbal	0.11
											negative verbal/total verbal	0.05
											qualified verbal	0.48
											qualified verbal/total verbal	0.42
											positive nonverbal	0.74
											English teachers	0.57[b]
											social studies teachers	0.64[b]
											positive nonverbal/total nonverbal	0.67
											negative nonverbal	-0.15
											negative nonverbal/total nonverbal	-0.09

[b]Spearman rank-order correlations. All others are Pearson product-moment correlations

[d]Refers to sum of all reinforcement categories

Studies of Teacher Behavior
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	DEPENDENT VARIABLE: No. of Observation Occasions	DEPENDENT VARIABLE: Type of Measure(s)	DEPENDENT VARIABLE: Behavior Observed	DEPENDENT VARIABLE: Stability Coefficients
Trinchero (1974) cont.	random	9th, 10th	similar	-	social studies, English	different	20	T	2	frequency count	Higher-order Questions:	
											teacher-initiated	0.16
											lower-order	0.17
											lower order / teacher initiated	-0.52
											higher-order	0.17
											higher order / teacher initiated	0.54
											Probing:	
											pupil-initiated talk	0.04
											opportunities to probe	0.24
											total probes	0.40
											total probes / opportunities to probe	0.35
											individual probes	0.31
											individual probes / total probes	0.20
											individual redirects	0.54
											individual redirects / total probes	-0.09
											group redirects	0.43
											group redirects / total probes	-0.11
											total repeating, rephrasing, refocusing	-0.09
											repeating, rephrasing, refocusing / total probes	-0.29
											total clarifying, prompting	0.62
											clarifying, prompting / total probes	0.03
											total critical awareness	-0.05
											critical awareness / total probes	0.14

Studies of Teacher Behavior
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	No. of Observation Occasions	Type of Measure(s)	DEPENDENT VARIABLE: Behavior Observed	Concept Lesson	Hypothesis Lesson	Value Lesson
												Stability Coefficients		
Ehman (1972)	-	9th	similar	-	social studies (historical/social science concepts, eliciting student hypotheses in response, analysis of value proposition or controversial issue)	same	84	U	2	frequency count	praises, encourages	0.41	0.47	0.42
											accepts, uses student ideas	0.20	0.03	0.47
											asks questions: overall	0.34	0.52	0.33
											cognitive-memory	-0.01	0.52	0.47
											convergent	0.12	0.13	0.01
											divergent	0.35	0.18	-0.01
											evaluative	0.32	0.21	0.23
											closed	0.26	0.07	0.26
											lectures	0.59	0.55	0.60
											gives directions	0.44	0.35	0.29
											criticizes	-0.08	0.25	-0.01
											productive silence/confusion	0.55	0.53	0.49
											unproductive silence/confusion	0.36	0.33	0.14
											I/D ratio	0.57	0.66	0.57

Studies of Teacher Behavior
(Continued)

	STUDENTS				CONTENT		TEACHER		DEPENDENT VARIABLE					
Study	Assignment to Teacher	Grade Level	Similarity	Type of Population	Type	Similarity	N	Level of Experience	No. of Observation Occasions	Type of Measure(s)	Behavior Observed	Stability Coefficients Sessions I-II	II-III	III-IV
Bellack (1966) (Ehman reanalysis)	non-random	10th, 12th	same	urban and suburban, mean verbal IQ = 70th percentile	social studies (economics)	similar	15	P	4	frequency count	teacher talk/student talk	0.84	0.79	0.41
											structuring	0.72	0.72	0.76
											soliciting	0.74	0.61	-0.03
											responding	0.66	0.32	0.22
											reacting	0.37	0.84	0.29

Studies of Teacher Behavior
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	DEPENDENT VARIABLE: No. of Observation Occasions	DEPENDENT VARIABLE: Type of Measure(s)	DEPENDENT VARIABLE: Behavior Observed	DEPENDENT VARIABLE: Stability Coefficients
Borg (1972)	non-random	-	same	-	varied	-	38	P	2	frequency count	teacher questions	0.14
											total zeros	0.18
											average length of pause	0.27
											words in pupil response to information-level question	0.01
											higher/fact order	0.07
											redirection	0.40
											prompting	0.30
											clarification	0.55
											repeating own question	0.23
											repeating pupil answer	0.22
											answering own question	0.19
											% of teacher talk	0.38

Studies of Teacher Behavior
(Continued)

	STUDENTS				CONTENT		TEACHER		DEPENDENT VARIABLE				
Study	Assignment to Teacher	Grade Level	Similarity	Type of Population	Type	Similarity	N	Level of Experience	No. of Observation Occasions	Type of Measure(s)	Behavior Observed	Stability Coefficients: Group Activities	Stability Coefficients: Individual Activities
Marzano (1973)	non-random	1st grade	same	-	language	different	27	P	5	ratings after every activity	Following the format: $\hat{\rho}^2$ teachers	0.52	0.54

Studies of Teacher Behavior
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	No. of Observation Occasions	Type of Measure(s)	Behavior Observed	Stability Coefficients
Wallen (1970)			same	-			33		10	global ratings	praise, encouragement	0.79
											warm, friendly	0.92
											control of personal behavior	0.72
											positive regarding children	0.83
											adaptable	0.71
											confident	0.91
											committed	0.86
											appropriate communication	0.81
											control of academic behavior	0.03
											personal enthusiasm	0.79
											sensitivity	0.83
										ratings every 5 minutes	Attention:	
											. procedures	0.01
											selling	-0.02
											compulsion	0.42
											questioning	0.19
											lecturing	0.19
											no attempt	0.67
										ratings every 5 minutes	Teacher Provisions for Pupil Response:	
											opportunities--	
											few	0.09
											many	0.24
											encouragement	0.20
											compulsion--	
											few	0.19
											many	0.12
											precludes--	
											few	0.38
											many	0.24

Studies of Teacher Behavior
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	No. of Observation Occasions	Type of Measure(s)	Behavior Observed	Stability Coefficients
Wallen (1970)	non-random	-	same	inner-city	-	-	33	P	10	ratings every 5 minutes	Individualization:	
											diagnosed	0.87
											unclear	-0.09
											not diagnosed	0.19
											no provision	0.14
											Grouping:	
											ability	0.62
											unclear	0.21
											not diagnosed	-0.04
										ratings every 5 minutes	Feedback:	
											from children	-0.22
											from procedures	0.00
											from teachers	0.43
											no provision	0.34

Studies of Teacher Behavior
(Continued)

Study	STUDENTS: Assignment to Teacher	STUDENTS: Grade Level	STUDENTS: Similarity	STUDENTS: Type of Population	CONTENT: Type	CONTENT: Similarity	TEACHER: N	TEACHER: Level of Experience	No. of Observation Occasions	Type of Measure(s)	Behavior Observed	Stability Coefficient
Wallen (1970) cont.	non-random	-	same	inner-city	-	-	33	P	10	ratings every 5 minutes	Classroom Management:	
											Response to Irrelevant Behavior:	
											hostility, reprimand	0.77
											direct control	0.61
											praise desired behavior	0.42
											humor, distraction	-0.13
											changing activity	-0.07
											empathy	0.09
											ignores	0.66
										ratings every 5 minutes	Classroom Management:	
											Response to Disruptive Behavior:	
											hostility, reprimand	0.14
											direct control	0.30
											praise of desired behavior	0.00
											humor, distraction	-0.01
											changing activity	0.00
											empathy	0.10
											ignores	0.25

TABLE 4
Clusters of Teacher Behavior Variables

Teacher Presentation

Variables (study*)	Stability Coefficients
teacher lecture (EP)	0.51
teacher presents (FP)	0.45
teacher informing vs. student statements (EP)	0.28
lectures (E)	0.55 to 0.60
lecturing (W)	0.19
teacher talk (M)	0.12 (SCIS) 0.76 (conventional)
teacher talk (B)	0.38
S/T ratio (C)	0.80
S/T ratio (Be)	0.41 to 0.84
I/D ratio (M)	0.18 (SCIS) 0.64 (conventional)
I/D ratio (C)	0.10
I/D ratio (E)	0.57 to 0.66

Presentation of Content

Variables (study*)	Stability Coefficients
explanation (EP)	0.16
descriptive interchange (EP)	0.18
optimum information amount (M)	0.37
higher cognitive level of student behavior (EP)	< 0.10
conceptual (EP)	< 0.10
problem-solving	0.27
knowledge structure cues (M)	0.29
association/drill (EP)	< 0.10
structuring (Be)	0.72 to 0.76
verbal fluency (M)	0.76
vagueness (M)	0.84
productive silence or confusion (E)	0.49 to 0.55
unproductive silence or confusion (E)	0.14 to 0.36

Teacher Questions

Variables (study*)	Stability Coefficients
teacher-initiated questions (T)	0.16
questioning (W)	0.19
teacher questions (B)	0.14
overall questions (E)	0.33 to 0.52
soliciting (Be)	-0.19 to 0.74
factual (M)	-0.12 (SCIS) 0.42 (conventional)
lower-order (T)	0.17
cognitive-memory (E)	-0.01 to 0.52
length of pupil response to information level questions (B)	0.01
relations (M)	0.05 (SCIS) 0.43 (conventional)
descriptive vs. inferential (EP)	< 0.10
observations (M)	0.05 (SCIS) -0.17 (conventional)
hypotheses (M)	0.19 (SCIS) 0.17 (conventional)
testing hypothesis (M)	0.00 (SCIS) 0.00 (conventional)
higher-order (T)	0.17
higher-order/fact (B)	0.07
convergent (EP)	0.14
convergent (E)	0.01 to 0.13
divergent (E)	-0.01 to 0.35
opinion (EP)	0.14
evaluative (E)	0.21 to 0.32
open (EP)	0.23
closed (E)	0.07 to 0.26
average length of pause (B)	0.27

Teacher Feedback

Variables (study*)	Stability Coefficients
feedback (W)	0.43
feedback (F)	0.63
total reinforcement (T)	0.64
provision for feedback (W)	0.34
reacting (Be)	0.29 to 0.54
support (F)	0.37
approval (F)	0.49
acknowledged (F)	0.61
positivity (F)	0.61
positive verbal reinforcement (T)	0.44
English teachers	0.04
social studies teachers	0.57
positive nonverbal reinforcement (T)	0.74
praises, encourages (E)	0.41 to 0.47
not evaluative (F)	0.58
neutrally reject (F)	0.48
qualified verbal reinforcement (T)	0.48
responding (Be)	0.22 to 0.66
criticism (F)	0.28
critical of pupil work (EP)	< 0.10
critical of wrong answers (EP)	0.16
self-reference question with negative answer (EP)	0.13
negative verbal reinforcement (T)	0.13
negative nonverbal reinforcement (T)	-0.15
criticizes (E)	-0.04 to 0.25

*Studies are identified as follows:

H = Hiller, et al. (1969)
EP = Emmer & Peck (1973)
C = Campbell (1972)
F = Friedman (1973)
M = Moon (1969, 1971)
T = Trinchero (1974)
E = Enman (1972)
Be = Bellack (1966)
B = Borg (1972)
Mz = Marzano (1973)
W = Wallen (1970)

Clusters of Teacher Behavior variables
(Continued)

Motivational Skills

Variables (study[e])	Stability Coefficients
interest (H)	0.20
selling (W)	-0.20
compulsion (W)	0.42
no attempt (W)	0.67

Expressive Teaching Style

Variables (study[e])	Stability Coefficients
expansive teaching behavior (EP)	0.22
enthusiasm (F)	0.53
encouragement I (F)	0.70
encouragement II (F)	0.73
teacher support (EP)	0.15

Student-Centered Teaching Style

Variables (study[e])	Stability Coefficients
feedback from students (W)	-0.22
student discussion (EP)	< 0.10
student vs. teacher ideas (EP)	-0.16
student presents vs. teacher questions (EP)	0.16
accepts, uses student ideas (E)	0.08, 0.20, 0.47
convergent vs. divergent teacher behavior (EP)	0.36

Interpersonal Behavior

Variables (study[e])	Stability Coefficients
student-teacher contacts (EP)	0.29
teacher seeks correct answer (EP)	0.26
teacher behavior and no student response (EP)	0.18
teacher behavior and incorrect student response (EP)	0.18
opportunity for few (W)	0.09
opportunity for many (W)	0.24
encouragement (W)	0.20
compulsion for few (W)	0.19
compulsion for many (W)	0.12
precludes for few (W)	0.38
precludes for many (W)	0.24
pupil-initiated talk (T)	0.04
opportunities to probe (T)	0.24
total probes (T)	0.40
individual probes (T)	0.31
individual redirect (T)	0.54
group redirect (T)	0.43
total redirections (T)	-0.09
total clarifying probes (T)	0.62
total critical awareness (T)	-0.05
redirection (B)	0.40
prompting (B)	0.30
clarification (B)	0.55
repeating own question (B)	0.23
repeating pupil answer (B)	0.22
answering own question (B)	0.19

[e]Studies are identified as follows:

H = Hiller, et al. (1969)
EP = Emmer & Peck (1973)
C = Campbell (1972)
F = Friedman (1973)
M = Moon (1969, 1971)
T = Trinchero (1974)
E = Ehman (1972)
Be = Bellack (1966)
B = Borg (1972)
Mz = Marzano (1973)
W = Wallen (1970)

Clusters of Teacher Behavior Variables (continued)

Affective Style

Variables (study[a])	Stability Coefficients
teacher affect (EP)	0.11
candor (EP)	0.19
I/d ratio (I)	-0.90

Following Procedures

Variables (study[a])	Stability Coefficients
following format (Mz)	
individual activities	0.52
group activities	0.54
attention-getting procedures (W)	0.01
feedback procedures (W)	0.03

Classroom Management

Variables (study[a])	Stability Coefficients
procedural (EP)	0.13
gives directions (E)	0.15, 0.38, 0.44
controlling (EP)	0.11
desist statements (EP)	0.39
criticizing, asking for behavior change (EP)	0.34
direct control (W):	
irrelevant behavior	0.61
disruptive behavior	0.30
hostility, reprimand (W):	
irrelevant behavior	0.77
disruptive behavior	0.17
praising desired behavior (W):	
irrelevant behavior	0.42
disruptive behavior	0.00
humor distraction (W):	
irrelevant behavior	-0.13
disruptive behavior	-0.01
empathy (W):	
irrelevant behavior	0.09
disruptive behavior	0.10
changing activity (W):	
irrelevant behavior	-0.07
disruptive behavior	0.00
ignoring (W)	
irrelevant behavior	0.66
disruptive behavior	0.25

Individualization

Variables (study[a])	Stability coefficients
individualization (W):	
diagnosed	0.87
unclear	-0.09
not diagnosed	0.19
ability grouping (W):	
diagnosed	0.62
unclear	0.21
not diagnosed	-0.04
no provision for individualization (W)	0.14

Global Ratings

Variables (study[a])	Stability coefficients
praise, encouragement (W)	0.79
warm, friendly (W)	0.92
control of personal behavior (W)	0.72
positive regarding children (W)	0.83
adaptable (W)	0.71
confident (W)	0.91
committed (W)	0.95
appropriate communication (W)	0.81
control of academic behavior (W)	0.03
personal enthusiasm (W)	0.79
sensitivity (W)	0.83

[a]Studies are identified as follows:

M = Miller, et al. (1959)
EP = Emmer & Peck (1973)
C = Campbell (1972)
F = Friesman (1973)
N = Moon (1968, 1971)
T = Trinthero (1974)
E = Emmer (1972)
Be = Bellack (1966)
B = Borg (1972)
Mz = Merzano (1973)
W = Wallen (1970)